DOCUMENT RESUME

ED 134 599 TM 005 995

AUTHOR	Brown, David Lile
TITLE	Faculty Ratings and Student Grades: A Large-Scale Multivariate Analysis by Course Sections.
PUB DATE	Dec 74
NOTE	127p.; Ph.D. Dissertation, University of Connecticut
AVAILABLE FROM	Dr. David Lile Brown, 36 Mansfield Hollow Road Extension, Mansfield Center, Connecticut 06250 ($5.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Bias; Effective Teaching; Factor Analysis; *Grades (Scholastic); *Higher Education; Multiple Regression Analysis; Predictor Variables; Rating Scales; *Statistical Analysis; *Student Evaluation of Teacher Performance; Validity
IDENTIFIERS	University of Connecticut

ABSTRACT

The purpose of this dissertation is to provide evidence bearing on the question of the influence of the grades students receive on their ratings of the college teachers who gave them those grades. Specifically, certain characteristics of the grade distribution within each course section are evaluated as predictors of the students' ratings of the teacher of that course section. Multivariate techniques were employed to evaluate data across an entire university. Over 30,000 anonymous student ratings of 2,360 course sections were collected after students had received final course grades, and without student or administrator knowledge that the ratings would be used in the study. Factor analysis was used to reduce the eight-item rating instrument to a single criterion variable. Subsequently, stepwise multiple regression analysis was used, both to reduce an initial battery of predictors to an optimally reduced subset, and to test the incremental importance of certain grading variables as predictors of the criterion.. The primary implication of the study is that there is a relationship between grades and ratings, but it only accounts for about nine percent of the variance in the ratings. There is a significant bias, but factors other than grades must also be influencing student ratings. Whether or not these other factors are valid measures of teaching effectiveness remains to be determined, but one seemingly invalid factor (the grading bias) has been identified. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED134599

FACULTY RATINGS AND STUDENT GRADES:
A LARGE-SCALE MULTIVARIATE ANALYSIS BY COURSE SECTIONS

BY

DAVID LILE BROWN

FIRST PUBLICATION, DECEMBER 1974

COPYRIGHT © DAVID LILE BROWN 1974.

TM005 995

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY
DAVID LILE BROWN
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ABSTRACT

FACULTY RATINGS AND STUDENT GRADES:
A LARGE-SCALE MULTIVARIATE ANALYSIS BY COURSE SECTIONS

David Lile Brown, Ph.D.
The University of Connecticut, 1974

Effective teaching has been a primary educational goal throughout history, and the evaluation of teaching is therefore one of the central concerns of education. Student ratings are often considered one of the best indicators of teaching effectiveness, because the student is in the most privileged position to view the teaching and to experience its effects. The use of such ratings has become widespread, but controversy rages over whether student ratings are valid measures of teaching effectiveness, especially when used for making decisions about faculty pay, promotion, and tenure.

It is possible that students are not mature enough to appraise teaching or, worse, that they may be exploiting rating systems to punish strict teachers and to reward lenient ones. If so, this would have important implications for the interpretation of student ratings. Proper interpretation requires an answer to the research question: What is the influence of the grades students receive on their ratings of the college teachers who gave them those

grades? This question has not been answered satisfactorily by previous research. The former evidence has been conflicting and inadequate in a number of ways. The present study, however, was intended to avoid many of the shortcomings of previous studies.

This study employed multivariate techniques to evaluate data across an entire university. Over 30,000 anonymous student ratings of 2,360 course sections were collected after students had received final course grades, and without student or administrator knowledge that the ratings would be used in this study. Factor analysis was used to reduce the eight-item rating instrument to a single criterion variable. Subsequently, stepwise multiple regression analyses were used, both to reduce an initial battery of predictors to an optimally reduced subset, and to test the incremental importance of certain grading variables as predictors of the criterion.

Results showed that the simple correlation between the average student grade in each course section and the average student rating of the teacher of that course section was .35, $\underline{p} < .000001$. Moreover, the average grade was the single best predictor, of those available, of the average rating; and when average grade was added to the optimally reduced subset of other predictors, it significantly improved the multiple correlation from .25 to .39, $\underline{F}\ (4, 2345) = 60.13$, $\underline{p} < .001$.

The primary implication of the results of this study is that there is a relationship between grades and ratings, but it only accounts for about 9% of the variance in ratings. In other words, there is a significant bias, but factors other than grades must also be influencing student ratings. Whether or not these other factors are valid measures of teaching effectiveness remains to be determined, but one seemingly invalid factor (the grading bias) has been identified through this study. The interpretation of student ratings should take this bias into account, and methods should be devised to eliminate it.

FACULTY RATINGS AND STUDENT GRADES:
A LARGE-SCALE MULTIVARIATE ANALYSIS BY COURSE SECTIONS

David Lile Brown
B.S., Tufts University, 1968
M.A., University of Connecticut, 1969

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

at

The University of Connecticut
1974

Copyright by

David Lile Brown

1974

APPROVAL PAGE

Doctor of Philosophy Dissertation

FACULTY RATINGS AND STUDENT GRADES:
A LARGE-SCALE MULTIVARIATE ANALYSIS BY COURSE SECTIONS

Presented by

David Lile Brown, B.S., M.A.

Major Advisor Ellis B. Page

Associate Advisor Steven V. Owen

Associate Advisor Robert K. Gable

Associate Advisor Marvin Rothstein

Associate Advisor Richard W. Whinfield

The University of Connecticut

1974

ACKNOWLEDGMENTS

This writer wishes to express his appreciation to:

Prof. Ellis B. Page, who, as his major advisor, has been a tremendous source of inspiration and encouragement during the past six years;

The other members of his advisory committee, Drs. Steven V. Owen, Robert K. Gable, Richard W. Whinfield, and Marvin Rothstein, whose suggestions, corrections, and insights were most valuable in guiding this effort to a successful conclusion;

The University of Connecticut Computer Center for the use of computer facilities;

Dr. Dorothy C. Goodwin, Mrs. Shirley Malinowski, Mrs. Althea McLaughlin, and Mrs. Meryl Kogan of the University of Connecticut Bureau of Institutional Research for their assistance with data collection and data editing;

Mr. Richard J. Stec and Miss Barbara J. Sochor of the University of Connecticut Student Data Systems and Programming Office for assistance with data collection and computer tape copying;

Mrs. Katherine J. Brown and Mr. Rudy Voit of the Registrar's Office for assistance with data collection; and

Lile H. and Helen W. Brown, his parents, for support, patience, and encouragement during this effort.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES v

LIST OF FIGURES vi

Chapter

I. STATEMENT AND DEFINITION OF THE PROBLEM 1
  Statement of the Problem 1
  Purpose of the Study 4
  Need for the Study 6
  Sources of the Data 18

II. SUMMARY OF RELATED RESEARCH 19
  Validity of Student Ratings 19
  Related Issues 25
  Summary 30

III. PROCEDURE 31
  Subjects 31
  Instrument 31
  Initial Predictor Variables 35
  Grading Variables 41
  Criterion Variables 42
  Statistical Analyses 43
  Summary 46

IV. RESULTS 47
  Factor Analyses of the Rating Instrument's Items 47
  Reduction of the Initial Battery of Predictor Variables 50
  Addition of the Five New Predictor Variables to the Optimally Reduced Subset 60
  Parallel Results Using Other Criteria 62
  Summary 65

V. DISCUSSION AND CONCLUSIONS 66
  Discussion of Results and Implications 66
  Conclusions and Recommendations 77

APPENDIX 81

BIBLIOGRAPHY 99

LIST OF TABLES

Table Page

1. Data for Figure 1, QPR Cutoff Points for Graduating Seniors, 1950-1974 13

2. Data for Figure 2, Median QPR's for All Undergraduates, 1952-1974 14

3. Data for Figure 3, Average of All A-Through-F Grades, 1961-1974 15

4. Mean Differences Between Quantitative and Verbal SAT Scores of Undergraduates with Different Majors 39

5. Factor Analyses of the Eight Items on the University of Connecticut Rating Scale for Instruction 48

6. Product-moment Correlations Among the Eight Items on the University of Connecticut Rating Scale for Instruction for 1972 and 1973 49

7. Means and Standard Deviations of the Predictor and Criterion Variables 51

8. Product-moment Correlations Among the Predictor and Criterion Variables 52

9. First Stepwise Multiple Regression Analysis—Reduction of the Initial Battery of Predictor Variables 59

10. Second Stepwise Multiple Regression Analysis—Addition of the Five Grading Variables to the Optimally Reduced Subset 61

LIST OF FIGURES

Figure Page

1. Quintile QPR cutoff points for graduating seniors, 1950-1974 10
2. Median QPR's of all undergraduates after the end of each semester, 1952-1974 11
3. Average of all A-through-F grades given, semester by semester, 1961-1974 12
4. The University of Connecticut Rating Scale for Instruction 32
5. Key to symbols used in Figures 6-18 85
6. Merge 1 86
7. First missing data input 87
8. Merge 2 88
9. Sort 1 89
10. Sort 2 90
11. Merge 3 91
12. Second missing data input 92
13. Merge 4 93
14. Merge 5 94
15. Third missing data input 95
16. Merge 6 96
17. Regression run 1 97
18. Regression run 2 98

CHAPTER I

STATEMENT AND DEFINITION OF THE PROBLEM

Statement of the Problem

Considerable recent controversy has centered around faculty evaluation methods, especially student ratings of college teachers (Centra, 1973; Doyle & Whitely, 1974; Frey, 1974; Zelby, 1974). According to several researchers (Bausell & Magoon, 1972a; Capozza, 1973; Carrier, Howard, & Miller, 1974; Centra, 1973; Costin, Greenough, & Menges, 1971; French-Lazovik, 1974; Menzie, 1973; Treffinger & Feldhusen, 1970), the use of student ratings to evaluate the faculty has become so widespread that it is now almost taken for granted at institutions of higher learning. Controversy rages, however, over the validity of such ratings, especially concerning "the trend toward formal, quantitative use of the results of the evaluations in determinations of faculty promotions and salaries" (Zelby, 1974, p. 1267).

Faculty reactions to student evaluation of teaching range from acclaim to outrage. Supporters point to the need for evaluation and to the potential that ratings may have for improving education. Opponents emphasize that students may not be proper judges, or that rating instruments may not ask students the right questions.

The whole issue is complex, but especially this question of validity. Are student ratings a valid measure of teacher effectiveness? The answer depends, of course, on definitions of validity and of teacher effectiveness. Since there are no universally accepted definitions of these, progress toward the development of such measures is impeded. Nevertheless, attempts at progress commonly are made with a rather vague notion of what "effective teaching" means.

If factual learning alone were the objective of effective teaching, one could validly measure various teachers' effectiveness through achievement testing. Indeed, many would support such measures as extremely valid, but others would insist that the "real goal in teaching is to impart philosophical values or to inculcate a special attitude toward learning rather than to simply help the student to master the subject matter" (Frey, 1974, p. 47). Still others would be inclined to consider the promotion of factual learning primary, but in combination with certain affective factors. How effective, for example, is the teacher in drawing students' attention and affection? Or how are the students made to feel about the specific subject matter they are learning? Does the teacher make it interesting?

Whatever view one takes of effective teaching, the major issue relevant to student ratings is still validity: Are student ratings a valid measure of effective teaching?

Some would argue that student ratings are valid because the student, as the consumer, is in the best position to view the teaching and to experience the effects of the teaching (see Centra, 1974; Costin, Greenough, & Menges, 1971; Frey, 1974; McKeachie, 1969).

Others would argue, for various reasons, that student ratings are not valid as a measure of teaching effectiveness. For example, LeComte (1974), Peck (1971), and St.Onge (1974) express strong doubts that students are in any position to judge the scholarship offered by the faculty. According to Costin et al. (1971), student ratings are typically challenged on the grounds "that student ratings are unreliable, that the ratings will favor an entertainer over the instructor who gets his material across effectively, that ratings are highly correlated with expected grades (a hard grader would thus get poor ratings) and that students are not competent judges of instruction since long-term benefits of a course may not be clear at the time it is rated" (p. 511).

One of the most important points of contention is whether or not students can be objective enough for their ratings to be valid. Even though student ratings are necessarily subjective opinions, they could still be valid measures of effective teaching if students were able to identify effective teaching (or at least some component or components thereof), and if their ratings were unbiased by irrelevant variables. Some believe, however, that students may bias their ratings in favor of lenient teachers, or at least in favor of the teachers in whose courses they get the highest grades. If so, this would have important implications for the interpretation of such ratings. It is even likely that such ratings would be at odds with larger educational objectives. For instance, "given a specific format, it is possible to adapt one's teaching technique to obtain a good or bad evaluation and . . . a good evaluation may be associated with a teaching technique of lesser educational value than a poor evaluation" (Zelby, 1974, p. 1268).

Therefore, one may pose the research question: What is the influence of the grades students receive on their ratings of the college teachers who gave them those grades?

Purpose of the Study

The purpose of this study is to provide evidence bearing on the research question. Specifically, certain characteristics of the grade distribution within each course section are evaluated as predictors of the students' ratings of the teacher of that course section. The following are the grading variables:

1. Average of the grades in the course section,
2. Standard deviation of the grades,
3. Variance of the grades,
4. Skewness of the grades, and
5. Kurtosis of the grades.

Stepwise multiple regression analysis is employed to evaluate these grading variables for their effectiveness as predictors, when used in combination with certain other predictors of such faculty ratings. These other predictors constitute an "optimally reduced subset" of the following variables:

1. Sex,
2. Appointment length,
3. Percentage employed,
4. Class size,
5. Number of years tenured,
6. Age,
7. Years since hiring,
8. Course level,
9. Title level,
10. Course location,
11. Department quantitativeness, and
12. Rate (percentage) of return of ratings.

An "optimally reduced subset" is defined as the subset which had the lowest standard error of estimate during a prior stepwise multiple regression analysis.

The hypothesis of this study is that the grading variables are strongly related to student ratings of college teachers, and that they will significantly improve the best prediction of those ratings attained by the other available predictors.

Need for the Study

The evaluation of teaching is one of the central concerns of education. Effective teaching has been a primary educational goal throughout history. As mentioned above, one commonly used measure of effective teaching is evaluation by the student. It would be hoped that student ratings would be valid and that they would promote effective teaching through feedback to teachers and administrators. However, it is not certain that students are mature enough, wise enough, or objective enough to evaluate such instructional behavior without bias. Are student ratings valid measures of effective teaching? Or are they more or less a "payoff" in a sort of game between teachers and students, in which teachers reward (or punish) students with grades, and students respond in kind with ratings?

The conflicting evidence to date has not satisfactorily answered this question. The extent and the importance of the relationship between grades and ratings have not been determined. Neither have any causes of such a relationship been positively identified, partially because the relationship itself has not been firmly established.

Former evidence on this question has been not only conflicting, but also inadequate in a number of ways. The results of previous studies will be summarized in chapter II, but most of these studies had one or more of the following potential shortcomings:

1. A relatively small sample size;
2. Sampling of only one or a few departments;
3. Sampling of only graduate assistant level teachers;
4. Sampling of team-taught courses;
5. Sampling of only freshmen level courses (or some other level);
6. Contamination by voluntarism (using only teachers who volunteer to be rated);
7. Contamination by sensitization (a new system is initiated, or a class is upset by a researcher suddenly appearing on the scene);
8. Contamination by knowledge of the study (students, faculty, or administrators know ahead of time that the ratings will be used in an "experiment");
9. Contamination by lack of anonymity protection for the students;
10. An indefinite treatment (students did not actually know what their final grade in the course would be at the time of the ratings); and
11. A lack of multivariate analysis (often only simple correlations among a few variables are reported, whereas multivariate techniques permit simultaneous examination of the influence of many variables and provide evidence of the "extent as well as the nature of any such influence" (Doyle & Whitely, 1974, p. 260)).

This present study does not suffer from any of the above potential shortcomings, and it should provide better evidence bearing on the research question than that previously found.

A further need for this study is institutional. While a somewhat similar study was conducted at the University of Connecticut ten years ago, no such study has been done recently. Changes have occurred in the rating instrument since that study of 10 years ago (Garber, 1964), and improvements have been made in the administrative procedures connected with the collection of ratings. There is reason to believe the data are less subject to collection errors now than formerly.

Furthermore, and perhaps more importantly, over the last 10 years, grades at the University of Connecticut have taken a sharp turn upward (see Figures 1-3 below). Whatever the relationship between grades and ratings may have been 10 years ago, that relationship well may have changed in light of the changes in grading, or in light of certain other changes related to the rise of student power in general. The advent, in the fall of 1968, of both pass-fail options and liberalized policies regarding the dropping of courses may have influenced the relationship between grades and ratings in some way. The appointment of students to formerly all faculty committees is another recent change which could possibly bear upon the research question.

Whether or not the trend in grading has had any effect on the relationship between grades and ratings, it is a striking trend and worthy of attention. Figure 1 shows the trend over the last 25 years of the quality point ratios (QPR's) of graduating seniors. The QPR is the grade average, computed as the sum of the quality points (number of credits in a course times the grade in that course, where "A" = 4, "B" = 3, "C" = 2, "D" = 1, and "F" = 0) divided by the total number of credits. At the University of Connecticut, QPR's are multiplied by 10 and range from 0 to 40, but for the sake of clarity to readers outside of the University of Connecticut, the range used throughout this study is the more prevalent 0 to 4. Figure 1 demonstrates that there was a great deal of consistency from 1950 until 1967 or 1968, but that the trend has been upward ever since. Table 1 provides the data used in Figure 1.

Figure 2 supplies further evidence of the trend in grading at the University of Connecticut, showing the median QPR's of all undergraduates after each semester. Note that each spring semester is represented by the year marks along the x-axis of Figure 2, while each fall semester is at the half-way point between year marks. The upward trend in median QPR's indicated by Figure 2 started about 1964, which is, as one would expect, three or four years before the start of the upward trend in graduating seniors' QPR's as shown in Figure 1. The data for Figure 2 are given in Table 2. Data for the spring semester of 1970 are missing because of a student strike near the end of that semester, following the U.S. bombing of Cambodia. In many courses that semester, final exams were cancelled and grades of "S" for

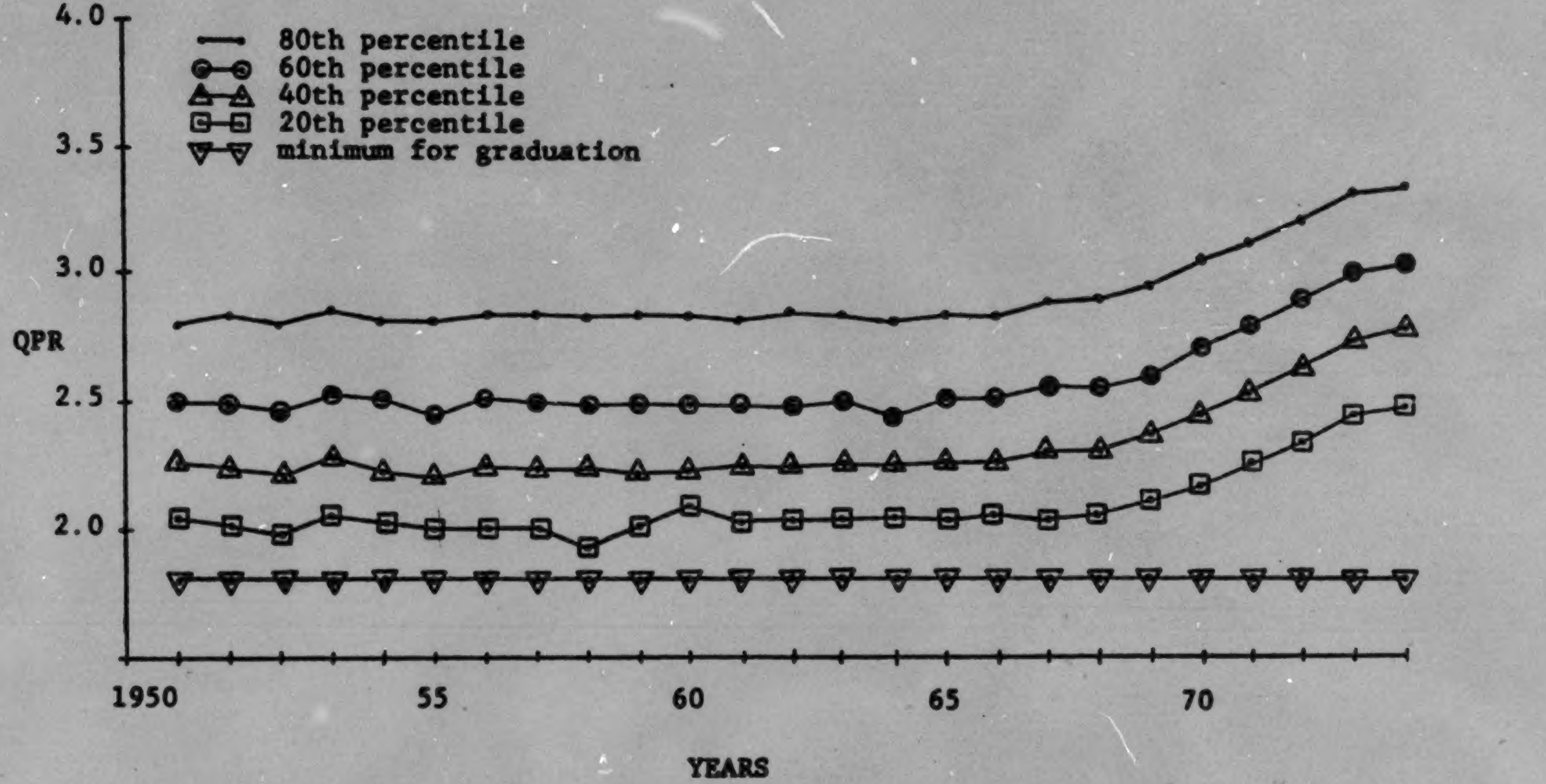

Figure 1. Quintile QPR cutoff points for graduating seniors, 1950-1974.

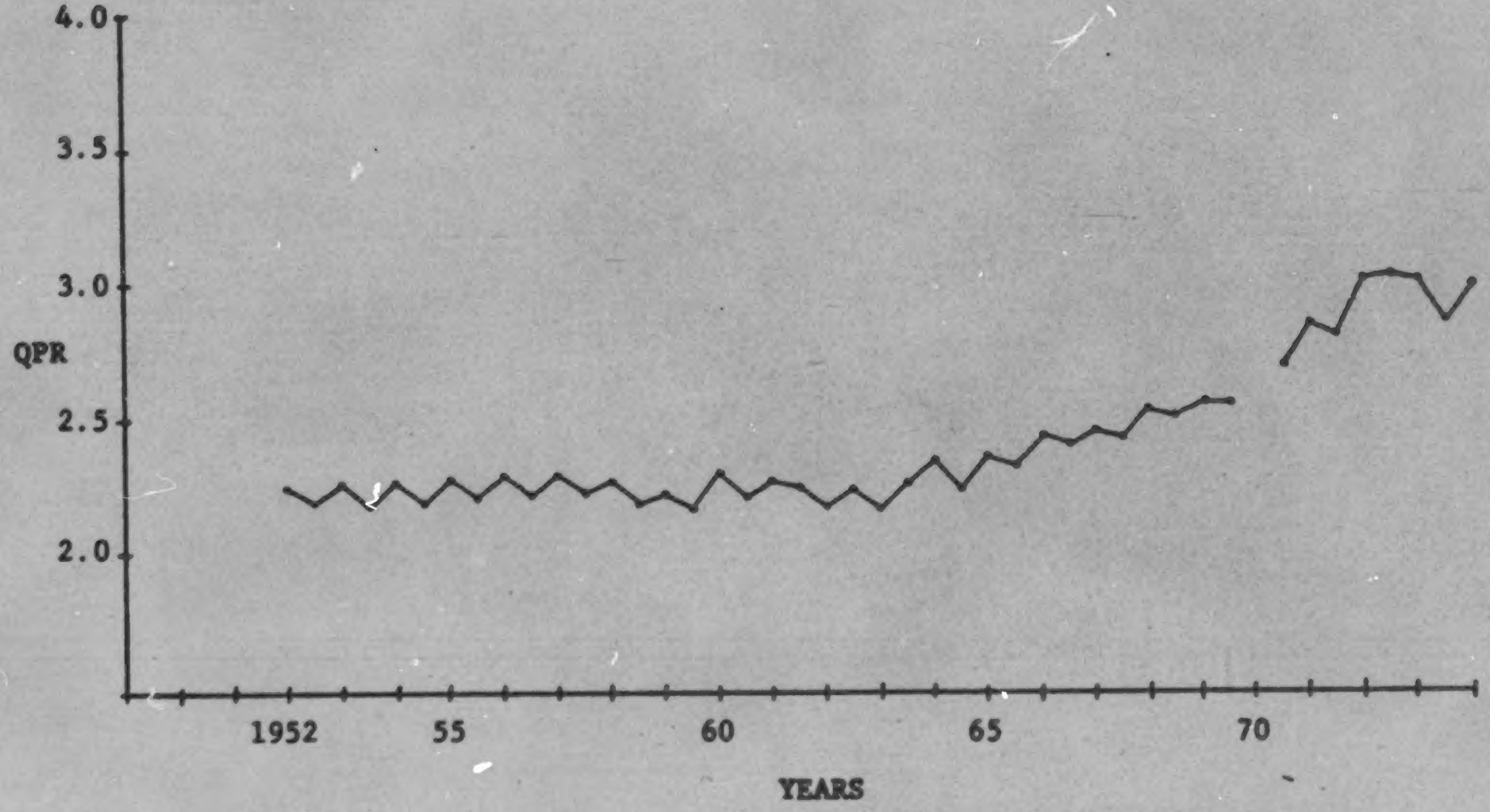

Figure 2. Median QPR's of all undergraduates after the end of each semester, 1952-1974.

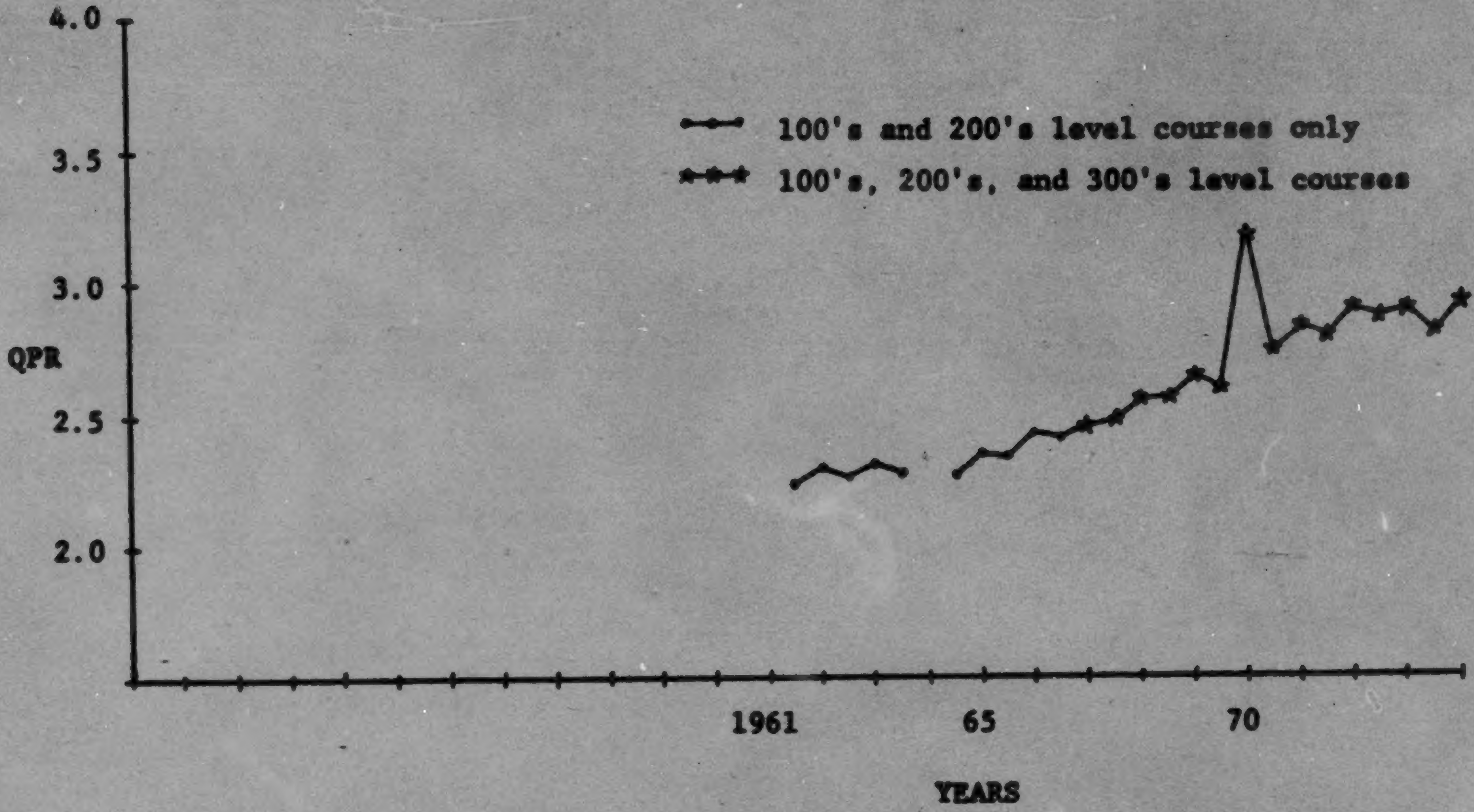

Figure 3. Average of all A-through-F grades given, semester by semester, 1961-1974.

Table 1

Data for Figure 1, QPR Cutoff Points for Graduating Seniors, 1950-1974

Graduating	Percentiles:			
Class of	80th	60th	40th	20th
1950	2.80	2.49	2.26	2.04
1951	2.83	2.49	2.24	2.02
1952	2.80	2.46	2.21	1.99
1953	2.85	2.52	2.28	2.06
1954	2.81	2.50	2.23	2.03
1955	2.81	2.44	2.20	2.00
1956	2.84	2.51	2.24	2.00
1957	2.84	2.49	2.23	2.00
1958	2.82	2.48	2.23	1.93
1959	2.83	2.48	2.21	2.00
1960	2.82	2.48	2.22	2.08
1961	2.81	2.48	2.24	2.03
1962	2.85	2.47	2.24	2.04
1963	2.83	2.49	2.25	2.04
1964	2.81	2.43	2.25	2.05
1965	2.83	2.50	2.26	2.04
1966	2.83	2.50	2.27	2.06
1967	2.88	2.55	2.30	2.03
1968	2.89	2.54	2.30	2.06
1969	2.94	2.59	2.36	2.10
1970	3.04	2.70	2.44	2.17
1971	3.11	2.79	2.52	2.25
1972	3.19	2.89	2.62	2.33
1973	3.30	2.99	2.72	2.43
1974	3.32	3.02	2.77	2.47

Note. The minimum QPR for graduation (0th percentile) was 1.80 every year.

Table 2

Data for Figure 2, Median QPR's for All Undergraduates, 1952-1974

Year	Spring Semester	Fall Semester
1952	2.27	2.21
1953	2.28	2.19
1954	2.27	2.21
1955	2.29	2.24
1956	2.31	2.24
1957	2.30	2.24
1958	2.28	2.20
1959	2.23	2.18
1960	2.31	2.23
1961	2.29	2.27
1962	2.19	2.25
1963	2.18	2.28
1964	2.35	2.26
1965	2.37	2.35
1966	2.45	2.43
1967	2.48	2.45
1968	2.56	2.53
1969	2.57	2.58
1970		2.72
1971	2.87	2.82
1972	3.02	3.04
1973	3.03	2.88
1974	3.02	

Note. Data for the spring semester of 1970 and the fall semester of 1974 (current semester) are missing.

Table 3

Data for Figure 3, Average of All A-Through-F Grades, 1961-1974

Year	Spring Semester	Fall Semester
1961		2.23
1962	2.29	2.26
1963	2.31	2.27
1964		2.26
1965	2.34	2.33
1966	2.42	2.40
1967	2.44	2.48
1968	2.56	2.55
1969	2.64	2.59
1970	3.18	2.73
1971	2.82	2.79
1972	2.89	2.86
1973	2.89	2.80
1974	2.92	

Note. Data for the following semesters are missing: spring of 1961, spring of 1964, and fall of 1974 (current semester).

"Satisfactory" were given. Some courses did have finals, and some teachers used regular grades with or without final exam scores. However, the median QPR for all undergraduates was not computed by the administration because of the drastic effect of the strike on grades. That effect is clearly indicated in Figure 3.

Figure 3 shows the trend in the average of the "A-through-F" grades given each semester. Again, the upward trend since 1964 is clearly indicated. The inclusion, starting with the spring semester of 1967, of 300's (graduate) level courses would account for some of the rise in this curve, but it should have resulted in a one-time jump. The constant upward trend can not be explained by the inclusion of these courses, which was unavoidable because of a change in administrative procedure. The data for Figure 3 are shown in Table 3.

There is no indication that the ability level of the students at the University of Connecticut has followed (or preceded) such a drastic trend as that indicated by Figures 1-3 (personal communications with the Admissions Office). A study by Baird and Feister (1972) found that a large number of faculties tended to give out the same distribution of grades over the years 1964-1968, even regardless of changes in the ability level of the students in some cases. The grading trend found at the University of Connecticut, however, is quite the opposite (higher grades without any such increase in student ability levels). This trend seems to be part of a nationwide trend toward leniency occurring since the period of the Baird and Feister study. Recent reports involving some 23 campuses across the nation ("Cal State," 1974; Ladas, 1974; "Too Many A's," 1974) have supplied evidence that a "grade glut has been spreading across academe" ("Too Many A's," 1974, p. 106) in the last few years.

It is not certain whether teachers are "simply being generous" or "are bribing students with good grades to get good grades themselves" ("Too Many A's," 1974, p. 106). But there is definitely an increased leniency on the part of this faculty, and it is possible that the relationship between grades and ratings is not the same as it was 10 years ago. Furthermore, it is likely that the upward trend in grading has not been universal across departments (cf. Ladas, 1974; Postman, 1974) nor across all teachers and that, as a result, the influence of grades, if any, on ratings would be more pronounced for some teachers and for some departments than for others.

Thus there are multiple reasons for conducting this study. The research question has not been satisfactorily answered; previous studies have suffered from several shortcomings; and changes in the ratings instrument and grading practices have occurred at this university (and apparently nationally). The present research is an attempt to provide better and more up-to-date knowledge concerning faculty evaluation by students.

Sources of the Data

This study uses data from the spring semester of 1973 at the University of Connecticut. Some of the data were obtained from admissions records or other branches of the university administration and were collected prior to 1973. The evaluation data were collected by the Bureau of Institutional Research in the customary manner established by that office and followed each spring. Detailed descriptions of the variables investigated and the rating instrument used appear in chapter III.

No one, including the students and the administrators who gathered the data, knew that the data would be used in this present study. Grade data were obtained from the registrar's records. Since ratings are done anonymously, it is not possible to match the individual students' ratings with their grades. Nevertheless, a great deal of information about the distribution of grades in each course section is known and can be matched with a large amount of other data about the course section and its teacher. Thus, the separate course sections (N = 2,360) were the units of analysis for this study. The loss of the ability to match grades and ratings for individual students is offset by certain compensations, such as the anonymity of the ratings data (which is what precludes the matching) and the ability to study data across an entire university.

CHAPTER II

SUMMARY OF RELATED RESEARCH

This chapter presents a summary of the theories and research which form the background to this study. The research related to the major issue of the validity of student ratings is covered first, followed by a summary of the research on several related issues.

Validity of Student Ratings

In 1971, Costin, Greenough, and Menges conducted a review of the literature on student ratings of college teachers. They stated that faculty members could judge the validity of student ratings "to the extent students' subjective criteria match the faculty members' goals in teaching" (p. 513), and thus, a determination must be made of "the basis on which students make their judgments" (p. 514). In their discussion of various possible bases for student judgments, Costin et al. cited conflicting evidence as to whether "students may judge instruction on the basis of its 'entertainment' value rather than on its information, contribution to learning, or long-term usefulness" (p. 517). They also stated that students may make rating judgments based on the grades received or expected in the course. This is the issue upon which this present study is focused.

Costin et al. reviewed 28 studies of the relationship between student ratings and grades. Of those, 15 found no significant relationship, 1 found a small negative correlation, and 12 found significant, but small, positive correlations. Typically the correlations did not exceed .30, but they could be considered collectively as evidence that the true relationship may be low and positive.

In an earlier study at the University of Connecticut, Garber (1964) used student ratings (responses to the eight separate items on the then current rating instrument) to predict "difference scores" (the differences between the students' grades and their "expected grades"). Current grade point averages were used as the expected grades. Garber obtained a multiple correlation of .43; correction for shrinkage yielded an $\underline{R}$ of .39 (a highly significant multiple $\underline{R}$, $\underline{p} < .001$). This outcome was for the multiple regression analysis using teachers as units of analysis. Using students as units of analysis, Garber obtained similar results ($\underline{R} = .31$; correction for shrinkage yielded an $\underline{R}$ of .28, $\underline{p} < .001$). He concluded that the two behaviors, student ratings of their teachers, and the direction in which the teachers tended to grade the students (higher or lower than their current grade point averages) were either directly or indirectly dependent on each other.

Bausell and Magoon (1972a), using a much larger sample than most other studies (N = 12,000 ratings, from which random stratified samples were drawn), "found strong, consistent biases in both instructor and course ratings which can be traced to (a) the grade the student expects to receive, and (b) the discrepancy between the student's expected grade and his GPA" (p. 1021). They concluded that student ratings of instructors are "axiomatically valid for their designed purpose, but must be interpreted with caution" (p. 1022) since the assignment of low grades may be proper in many cases, but would result in lower ratings. The implication is that administrative use of ratings for pay, promotion, and tenure decisions may require great caution.

A study by Treffinger and Feldhusen (1970) used characteristics of the students to predict student ratings of their teachers. They concluded that student variables, including grades received, were not significant predictors. They found that generalized pre-course ratings were most often the best predictor of end-of-course ratings, and therefore they concluded that the "student's rating of the course is clearly a complex interaction of his initial feelings, certain cognitive and affective characteristics of teachers and pupils, and instructor performance" (p. 622). Treffinger and Feldhusen noted that "it should be the instructor's performance or ability as a teacher and students' reliable perceptions and evaluations of the performance which constitute the majority of the variance in instructor ratings" (p. 622). They suggested that the difference between pre- and post-course ratings might be more useful than post-course ratings. But other research (especially Holmes, 1972) indicates that such a difference may be very heavily affected by changes in grades from initial expected grades to end-of-course expected or actual grades.

Bausell and Magoon (1972b) found that "rating changes did occur as a function of changes in grade expectancy" (p. 10). Holmes (1972) conducted an experiment in which expected grades were "disconfirmed." "Half of the students who deserved and expected A's or B's were given their expected grades, while half were given a grade one step lower than expected" (p. 130). Correct grades were given after ratings were collected. He concluded that "if students' grades disconfirm their expectancies, the students will tend to deprecate the instructor's teaching performance in areas other than his grading system" (p. 130, emphasis added). These results support a theory that students are not objective; that they use ratings as a "payoff;" and that ratings might therefore be invalid.

Many studies have investigated the relationship between ratings and achievement rather than grades. Such studies are, of course, directly related to the validity issue since it is obvious that high achievement constitutes a primary goal of effective teaching. If students rate highest the teachers from whom they learn the most, then student ratings would have a certain validity. Furthermore, such studies often provide important insights into the relationship between ratings and grades insofar as achievement and grades are closely interrelated.

One such study (Rodin and Rodin, 1972) stirred a great deal of interest in the relationship between ratings and grades. The authors found a significant negative correlation between amount learned and ratings of the teachers (teaching assistants in one large undergraduate calculus course with 12 sections). Controversy was stirred over their conclusion that "perhaps students resent instructors who force them to work too hard and to learn more than they wish" (p. 1166).

Another investigator (Capozza, 1973) agreed. He concluded that "the hypothesis that students give good ratings to classes in which they learn a great deal must be rejected. Emotional factors such as grades and perhaps a distaste for the hardships of learning seem to bias the results in the opposite direction" (p. 127). Though his sample (250 students in eight course sections) was larger than many, Capozza himself considered it small.

Gessner (1973) found results opposite those of the Rodins and Capozza. He criticized the Rodin and Rodin study for its use of teaching assistants, whose instructional role was apparently an ancillary one. Gessner's own data, however, were based on only one course (*N* = 78), and it was team-taught. There can be no generalization, then, across courses or instructors.

A study by Potter, Nalin, and Lewandowski (1973) employed a better research design than others (with students and teachers randomly assigned to classes). Unfortunately, however, teacher trainees were used. They found a modest, but significant, correlation between ratings and achievement when the effect of aptitude was partialled out. Although their sample was small ($\underline{N}$ = 254 students in 24 course sections), they observed that the magnitude of the correlation between achievement and ratings seemed to be lower when the correlation between achievement and aptitude was higher, and vice versa. They concluded: "the stronger the relation between aptitude and achievement, the less the relation between achievement and rating" (p. 2). A corollary of this conclusion might be that ratings may suffer from a grading bias in direct proportion to how arbitrary the grading is (cf. Holmes, 1972).

There are two conflicting explanations often proposed for a negative correlation between achievement and ratings: (a) better learners are more critical of teachers or (b) there is resentment of the hard work teachers force students to do in order to achieve. Rodin (1974), however, indicated that we may not need either of these theoretical explanations, since the true relationship well may be positive. She reviewed several studies and found that most of the correlations between ratings and amount learned tended "to lie in the range $\underline{r}$ = .20 to $\underline{r}$ = .30 . . . which indicates that amount learned accounts for only about 4 to 9 percent of the variance in student ratings" (p. 5). Rodin concluded that student ratings are not indicative of amount learned. She compared student ratings of teaching to consumer ratings of the palatability of peanut butter, and argued that, just because the consumer ratings may be poorly related to laboratory ratings of nutritiveness, the validity of palatability ratings is not thereby impugned (unless one expects the palatability ratings to tell us all about peanut butter). Rodin implied, therefore, that student ratings may be a valid measure of one component of effective teaching, and that grades could be one important influence on that component.

Related Issues

Several issues, which have been hinted at above, are directly related to validity. One is the possible presence of a "halo-effect." That is, are global impressions at work, masking the separate traits of effective teaching? Do the individual items on rating instruments provide useful information, or do they form large clusters of traits? Such clusters of traits could be valid measures of teaching effectiveness even if the separate items or traits were not.

Strong halo effects were noted by Royce (1960), Garber (1964), Potter, Nalin, and Lewandowski (1973), and Widlak, McDaniel, and Feldhusen (1973). Hoyt (1969) found a halo effect in student ratings of courses and concluded that it interfered with students' ability to discriminate fairly among the traits of the course covered by the rating instrument. Costin, Greenough, and Menges (1971), however, cited several studies in support of a theory that ratings on multiple attributes of instruction are low in halo effect.

Widlak et al. reviewed a representative sample of 22 studies over the past 30 years. They concluded that, while most rating instruments can be factored into at least two components, there is likely to be a halo effect "so strong . . . that the specific item ratings may have little diagnostic value in assessing a teacher's strengths and weaknesses and, ultimately, low potential for improving teaching" (p. 10). Widlak et al. found three factors throughout the studies they reviewed. They called these the "Actor," "Interactor," and "Director" factors (referring to roles of teaching). These might also be called "Performance," "Rapport with students," and "Course structure/difficulty" factors.

Another issue, somewhat related to the halo-effect issue is the persistence of first impressions or preconceptions. The findings of Bausell and Magoon (1972b) suggest that first impressions and preconceptions are persistent. They found that ratings of teachers after the first day of classes were very durable, and that they were highly correlated ($\underline{r}$ = .67, $\underline{n}$ = 20 courses) with ratings taken at the end of the course.

As mentioned above, Treffinger and Feldhusen (1970) found that generalized pre-course ratings were most often the best predictor of end-of-course ratings. Furthermore, Parent (1971) suggests that ratings should be taken early in the course to provide feedback to the teachers in time to modify the teaching performance before the courses were finished and "wasted." If early attitudes are persistent, then end-of-course ratings would not improve, but Parent suggests that end-of-course ratings should be eliminated. If ratings were always done very early in the course, some of the grading bias (if there were one) might be eliminated. Some bias might remain, however, since expected grades could still have an influence, as could grades on various quizzes and papers.

On the one hand, the durability of first impressions might indicate that student ratings are invalidated, because they are not measures of teaching effectiveness over the entire course. On the other hand, it may be that, whatever student ratings are measuring, it simply does not take very long to evaluate it (especially, perhaps, the Actor factor). In the case of preconceptions, it could be that there is accurate foreknowledge about the course or instructor from word-of-mouth advice from other students, or from prior experiences with the instructor himself or with similar courses taught by others.

Costin et al. suggested that student ratings might be shown to be valid if there were a high correlation between student and peer ratings (agreement between different judges). The results of the Bausell and Magoon (1972b) study suggest that ratings by outside observers, including other faculty members, would tend to correlate highly with student ratings, even though the outside observers might view only a few class sessions. Touq and Feldhusen (1973) supported this view, as do the results of four studies (r's from .30 to .63) cited by Costin et al. (1971). Centra (1974), however, found that while there is agreement between student and peer ratings of teachers, the peer ratings were "less reliable" (low interjudge agreement among the peers). He concluded that both student and peer ratings should be used, but to measure different traits. Costin et al. cautioned that peer ratings may be influenced by student ratings through hearsay (see also Jaeger & Freijo, 1974). It is possible, however, that a grading bias is responsible for the difference between student and peer ratings, or at least some portion of the difference. One implication is that peer ratings might be used to partial out the grading bias if one exists.

Still another issue relates to the definition and measurement of effective teaching. It is the issue over whether or not the good researcher is likely to be a good teacher. Stallings and Singhal (1969) found a significant correlation between teacher evaluations and research output (measured by weighted combinations of the number of books, articles, etc.). They concluded, as many others have, that productive researchers are usually very good teachers, but that students often have the stereotype that very active researchers neglect their teaching. If pervasive, such a stereotype would tend to lower the validity of student ratings of researchers, since the ratings might be based on a misconception. McDaniel and Feldhusen (1970) found that students did bias their ratings against heavy researchers or prolific writers; they theorize that the students may be correct — that researchers do neglect their teaching.

Many studies (notably Aleamoni, 1974; Bendig, 1952; and Elmore and LaPointe, 1974) have examined relationships between faculty ratings and certain characteristics of the teachers or courses. There was, according to them, little consensus about the characteristics of the "best" or "worst" teachers or classes. At the University of Connecticut, however, administrative findings over the past few years indicate there are definite differences in ratings across several variables. Historically, students at this university rate male teachers slightly higher than female teachers overall, but not on every item of the scale. In general, the smaller the class size, the higher the ratings of the instructor have been. Similarly, classes at the branches have been rated higher than courses at the main campus at Storrs, Connecticut. Officials have noted, however, that this last difference may be due to the fact that the largest course sections are taught at Storrs.

Historical evidence also indicates that at the

University of Connecticut, the more advanced the course, the higher the ratings. Tenured faculty have received better ratings than non-tenured teachers. With respect to age, a teacher's ratings seem to peak during the 31-to-40 age period, with ratings rising steadily to that point and then declining only slightly and very slowly afterwards. Thus, variables such as age, sex, tenure status, and class size might be good predictors of student ratings.

Summary

A careful scrutiny of the literature reveals that very few multivariate studies have been conducted in the area of this problem, and that several criticisms (mentioned in chapter I) may be leveled at most of the simpler, but more numerous, correlational studies. Furthermore, there is disagreement among the researchers, and directly conflicting results from their studies have been confounded by the use of many different research methods, different rating instruments, and different sample characteristics (including different units of analysis). In sum, the research question has not been answered satisfactorily, though several interesting concepts and theories have emerged which could possibly prove helpful, once better evidence is obtained.

CHAPTER III

PROCEDURE

This chapter describes the subjects, the rating instrument, the predictor variables, the criterion variables, and the statistical methods used in this study. Appearing in the appendix are macro flowcharts of the computer programming steps required to assemble the data for analysis.

Subjects

The "subjects" in this study should be considered to be the 2,360 course sections for which complete data could be assembled. Since only 89 course sections were deleted for missing data, this study covered 96.37% of the entire population of 2,449 course sections evaluated following the spring semester of 1973 at the University of Connecticut. Over 30,000 rating forms were returned, representing about 55% of the students who were sent them.

Instrument

The University of Connecticut Rating Scale for Instruction (UCRSI; Bureau of Institutional Research, University of Connecticut, 1971) is presented in Figure 4.

Figure 4. **UNIVERSITY OF CONNECTICUT RATING SCALE FOR INSTRUCTION** FORM NO. AA 118 11/71

INSTRUCTOR ______________ COURSE ______ DEPT. ______ SEC. ______ ENRL.

1. Knowledge Of Subject	1☐ 2☐ 3☐ Knowledge of field inadequate	4☐ 5☐ 6☐ 7☐ Reasonably knowledgeable	8☐ 9☐ 10☐ Highly expert
2. Presentation Of Material	1☐ 2☐ 3☐ Very hard to follow	4☐ 5☐ 6☐ 7☐ Reasonably understandable	8☐ 9☐ 10☐ Makes subject very clear
3. Balance of Breadth & Detail	1☐ 2☐ 3☐ Gets bogged down in trivia	4☐ 5☐ 6☐ 7☐ Reasonable balance	8☐ 9☐ 10☐ Good balance of breadth & detail
4. Enthusiasm for Subject	1☐ 2☐ 3☐ Seems irksome to him	4☐ 5☐ 6☐ 7☐ Mildly interested	8☐ 9☐ 10☐ Displays great enthusiasm
5. Fairness in Marking	1☐ 2☐ 3☐ Partial and prejudiced	4☐ 5☐ 6☐ 7☐ Reasonably fair	8☐ 9☐ 10☐ Very fair and impartial
6. Attitude Toward Student	1☐ 2☐ 3☐ Unsympathetic and intolerant	4☐ 5☐ 6☐ 7☐ Reasonably good; average	8☐ 9☐ 10☐ Sympathetic and understanding
7. Personal Mannerisms	1☐ 2☐ 3☐ Distracting and irritating	4☐ 5☐ 6☐ 7☐ Minor peculiarities; average	8☐ 9☐ 10☐ Completely acceptable
8. Over-All Summary As A Teacher	1☐ 2☐ 3☐ Unsatisfactory	4☐ 5☐ 6☐ 7☐ Average	8☐ 9☐ 10☐ Outstanding

MAKE NO MARKS IN THIS SPACE

FACULTY EMPLOYEE NUMBER

INSTRUCTOR ______________ COURSE ______ DEPT. ______ SEC. ______ ENRL.

1. Knowledge Of Subject	1☐ 2☐ 3☐ Knowledge of field inadequate	4☐ 5☐ 6☐ 7☐ Reasonably knowledgeable	8☐ 9☐ 10☐ Highly expert
2. Presentation Of Material	1☐ 2☐ 3☐ Very hard to follow	4☐ 5☐ 6☐ 7☐ Reasonably understandable	8☐ 9☐ 10☐ Makes subject very clear
3. Balance of Breadth & Detail	1☐ 2☐ 3☐ Gets bogged down in trivia	4☐ 5☐ 6☐ 7☐ Reasonable balance	8☐ 9☐ 10☐ Good balance of breadth & detail
4. Enthusiasm for Subject	1☐ 2☐ 3☐ Seems irksome to him	4☐ 5☐ 6☐ 7☐ Mildly interested	8☐ 9☐ 10☐ Displays great enthusiasm
5. Fairness in Marking	1☐ 2☐ 3☐ Partial and prejudiced	4☐ 5☐ 6☐ 7☐ Reasonably fair	8☐ 9☐ 10☐ Very fair and impartial
6. Attitude Toward Student	1☐ 2☐ 3☐ Unsympathetic and intolerant	4☐ 5☐ 6☐ 7☐ Reasonably good; average	8☐ 9☐ 10☐ Sympathetic and understanding
7. Personal Mannerisms	1☐ 2☐ 3☐ Distracting and irritating	4☐ 5☐ 6☐ 7☐ Minor peculiarities; average	8☐ 9☐ 10☐ Completely acceptable
8. Over-All Summary As A Teacher	1☐ 2☐ 3☐ Unsatisfactory	4☐ 5☐ 6☐ 7☐ Average	8☐ 9☐ 10☐ Outstanding

MAKE NO MARKS IN THIS SPACE

THIS RATING IS TO BE ENTIRELY IMPERSONAL. DO NOT MAKE ANY MARKS
ON THIS PAGE WHICH COULD SERVE TO IDENTIFY YOU.

DC 6717

NORTHEAST BUSINESS FORMS
MERIDEN, CONN. 06450

The UCRSI is the instrument currently used at the university in a program of faculty evaluation that had its origin in the late 1940's. It started with a University Senate concern with the quality of teaching. It was felt that there might be an overemphasis on the "publish or perish" ethic.

At first, the university evaluated only certain "target" teachers (those for whom the administration requested information, usually for career decisions). But by the late 1960's, the process had grown into one that covered every faculty member teaching a "ratable" course section. The nonratable course sections are those deemed not ratable in the normal way, such as seminars, independent study courses, field work, practice teaching, team-taught courses, graduate assistant-taught courses, and others. The decision as to whether or not a specific course section is ratable is made by the individual department chairmen each year.

After students have received their grades for the spring semester, and after they have moved to their summer addresses, the rating forms are mailed to the students with their teachers' names and their course sections already filled in (see Figure 4). The university pays for postage both ways, and the anonymity of the students is guaranteed since students' names or identification numbers appear nowhere on returned forms. The rate of return of ratings has remained around 55% over the recent years.

Ratings are done only after each spring semester. This is done for two reasons: (1) so that students will fill out rating forms at their homes in the summer away from the influences of their fellow students, and (2) especially so that the cost of evaluation will not be doubled (by repeating the evaluations each semester).

Results are machine tabulated and sent both to the individual teachers and to their department chairmen. Cumulative average scores (cumulative since 1969) accompany each teacher's ratings for the recently ended course sections. Since the Bureau of Institutional Research edits each of the incoming rating forms by hand (including checking for the clarity of markings), and since invalid forms are removed, high accuracy is claimed for the ratings data.

The ratings data are provided to promotion and tenure committees by the department chairmen, along with other pertinent information. There has been some concern at the university that someone might attribute significance to very small differences in ratings, when in fact, only extreme differences in ratings have any practical meaning. There is also some concern about the validity of the UCRSI. There is no certainty as to what it measures, although university officials have noted that the ratings seem to have a moderately high "reliability." That is, large samples of student ratings show a high degree of agreement with each other. What the students are agreeing about, however, is not absolutely clear.

Initial Predictor Variables

The initial battery of predictors of student ratings of college teachers included twelve variables. Presented here are brief descriptions of each of these predictors and of how each was derived.

Sex. The sex of each teacher was obtained from his or her professional history, filed with the administration. No cases were lost for missing data for this variable. The characters M and F, for male and female, were converted to values 1 and 2 respectively before computation began.

Appointment length. The number of months of each teacher's appointment was also obtained from the professional history. In most cases, teachers are appointed each year for a term of 9 months. A few have 10 month appointments, and several are for 11 months. The values 9, 10, and 11 were used for this variable, and no cases were lost for missing data.

Percentage employed. Most teachers at the University of Connecticut are employed full-time, regardless of their appointment length. Several are part-time, however, and the percentages vary. Values from 1 to 100, representing percentages of full-time employment, were obtained from the professional history, and no cases were lost for missing data.

Class size. The number of students enrolled in each course section at the end of the semester was obtained from the grade distribution data. It equals the total number of all grades given, including incompletes, etc. No cases were lost for missing data for this variable.

Number of years tenured. The year that a teacher was tenured (if he was) was obtained from the professional file. The number of years since being tenured was computed (zero if not tenured) by subtracting the tenure year from 1973. No data were missing.

Age. The birthdate of each teacher was obtained from the professional history file, and each teacher's chronological age in completed years was computed as of May 8, 1973 (the end of the spring semester).

Years employed at the University of Connecticut. The hiring date of each teacher was also obtained from the professional history. The number of years of continuous service at the university was computed as of May 8, 1973 (the end of the spring semester) for each case. None of the cases were lost on account of missing data for this variable.

Course level. Course numbers were obtained from the rating responses records. The level of the course was then computed by dividing the course number by 100 and dropping the digits to the right of the decimal point. The resulting values (0, 1, 2, 3, and 4) represent courses with numbers 1-99, 100-199, 200-299, 300-399, and 400-499 respectively. No cases were lost for missing data for this variable.

Title level. Each teacher's classification code was obtained from the professional file. A list of the classification codes belonging to each title level (Instructor, Assistant Professor, Associate Professor, and Professor) was created with the aid of personnel from the Bureau of Institutional Research for use in one of the merging computer programs. Title levels (values 1, 2, 3, and 4 for Instructor, Assistant Professor, Associate Professor, and Professor respectively) were assigned to each teacher according to his or her classification code. Two course sections were lost since one teacher's classification code could not be correctly grouped with any one level.

Course location. A branch code was obtained for each course section from the rating responses data. A value of 1 was assigned to the course location if the course was taught at any of the branches (Hartford, Stamford, Southeastern, or Hartford M.B.A.), and a value of 2 was assigned to course sections taught at the Storrs main campus. No cases were lost for missing data for this variable.

Department quantitativeness. Student records for juniors and seniors (majors are not declared earlier) included the major department and his two Scholastic Aptitude Test Scores (Verbal and Quantitative). For each department in which students declared a major, the average quantitativeness ($\underline{DQ}_j$) was computed as the average difference in Quantitative and Verbal SAT scores of the students majoring in that department (j). The formula used to compute the average quantitativeness for each department was the following:

$$\underline{DQ}_j = \Sigma(\underline{SATQ}_{ij} - \underline{SATV}_{ij}) / \underline{n}_j \qquad (1)$$

where, in the jth department, $\underline{SATQ}_{ij}$ is the ith student's Quantitative SAT score, $\underline{SATV}_{ij}$ is the ith student's Verbal SAT score, and $\underline{n}_j$ is the number of students majoring in that department. The results of these computations are presented in Table 4. Since some departments had very few students as majors, the quantitativeness figures for those departments should be interpreted cautiously. With small $\underline{n}$'s, those departments may have radically different quantitativeness figures from one year to the next.

Note also that departments high in quantitativeness have high positive values (maximum of 169.5), and departments low in quantitativeness have low or negative values (minimum of -135.0). Being very high or very low (or anywhere between for that matter) on this scale of quantitativeness does not signify anything about the quality of the department or its *average* SAT scores. One could argue either way that it is better to be highly "verbal" or highly "quantitative." Yet for suggesting a profile, the scale appears to be very meaningful; it has high face validity.

There are 10 departments in which no undergraduate students may major (Aerospace R.O.T.C., Biobehavioral Science, (general) Engineering, Interdepartmental,

Table 4

Mean Differences Between Quantitative and Verbal SAT Scores of Undergraduates with Different Majors

Department			
Name	Mean	n	S.D.
Statistics	169.50	4	49.10
Civil Engineering	119.99	101	73.10
Mathematics	106.04	144	87.04
Chemical Engineering	105.42	26	80.99
Mechanical Engineering	95.89	35	71.18
Electrical Engineering	94.64	78	96.26
Accounting	82.23	140	81.77
Pharmacy	80.93	45	84.94
Finance	71.60	102	84.64
Business	68.06	190	87.95
Physics	64.68	19	70.33
Italian	64.50	2	46.50
Geography	62.25	12	70.35
Physical Education	61.47	87	72.46
Geology	60.83	24	87.42
Chemistry	59.00	43	82.62
Agricultural Engineering	58.50	8	84.80
Industrial Administration	57.55	60	86.65
Agricultural Economics	56.33	3	42.87
Marketing	52.44	94	81.47
Pre-Veterinary	49.00	13	89.70
Nutritional Science	43.00	7	76.93
Biology	41.86	345	92.16
Horticulture	41.07	99	82.83
Animal Industries	39.98	49	85.01
Economics	31.99	101	77.50

(Continued on the next page)

Table 4 (Continued)

Department			
Name	Mean	n	S.D.
Agriculture and Natural Resources	29.00	1	0
Spanish	22.82	38	94.36
Physical Therapy	19.25	181	90.33
Political Science	18.85	189	84.80
Foods & Nutrition	17.86	22	69.38
Family Economics	17.57	7	76.56
German	16.57	14	119.31
Clothing, Textiles, & Interior Design	15.79	62	97.21
Speech	14.66	110	88.50
Child Development & Family Relations	14.10	174	81.78
Psychology	13.49	321	93.86
History	13.37	188	94.02
Music	10.56	45	93.50
Medical Technology	10.29	35	84.86
Sociology	10.16	225	89.27
Education	9.60	354	81.45
Art	6.11	82	77.10
Philosophy	-3.46	33	79.87
Nursing	-4.24	193	76.47
Anthropology	-5.30	57	86.03
French	-17.50	47	73.60
Dramatic Arts	-23.02	47	104.45
Russian	-29.71	7	127.18
English	-33.09	401	88.84
Classics	-135.00	1	0

Journalism, Linguistics, Metallurgy, Polish, Portuguese, and Science). The 56 course sections that were taught in these 10 departments were excluded from this study because of missing data for department quantitativeness.

Rate of return of ratings. For each course section, the percentage of rating forms returned by the students was computed. The number of ratings returned (obtained from the ratings data) was divided by the enrollment in the class (see class size). Values for rates of return were permitted to range from .001 to 1.000, and no cases were dropped on account of missing data.

Grading Variables

Average grade. The number of each type of grade given in each course section was obtained from the grade distribution data (from the registrar's records). Grades such as "I" for Incomplete and "P" for Pass were ignored, and the average of the "A-through-F" grades was computed for each course section (values permitted from 0.00 to 4.00). There were 27 cases deleted on account of missing data for this variable. In these 27 course sections, no grades were given in the "A-through-F" range. Note that these course sections differ from those in which all "F's" were given (resulting in an average grade of 0.00), and thus the 27 cases could not logically be included in the study with any particular value for an average grade.

Standard deviation of grades, Variance of grades, Skewness of grades, and Kurtosis of grades. These characteristics of the distribution of the "A-through-F" grades in each course section were computed at the same time as the average grade, using the registrar's grade distribution data. No further cases (beyond the 27 cases lost for average grade missing data) were lost for missing data for these variables.

Criterion Variables

Ten potential criterion variables were computed for each course section using the rating responses data. The process of selecting the criteria used in the stepwise multiple regression analyses is explained in the next section and in the next chapter. Descriptions of the 10 potential criteria and how they were derived are presented here.

Items 1-8. The average rating on each of the eight items on the University of Connecticut Rating Scale for Instruction (UCRSI, see Figure 4) was computed for each course section. Four course sections (in which no one answered Item 8) had to be deleted for missing data.

Average of Items 1 through 8. The average of the responses to all of the items on the UCRSI was computed for each course section using the ratings data. No cases were dropped on account of missing data for this variable (none beyond the four dropped for Item 8 missing data).

Average of Items 1 through 7. The average of the first seven items on the UCRSI was also computed from the rating responses data, and there were no missing data.

Statistical Analyses

In order to determine the number and nature of components or dimensions underlying the rating instrument items, two principal components factor analyses (of 1972 and 1973 ratings data) were conducted prior to the main effort of this study. The results of these (details presented in chapter IV) were instrumental when it came to making decisions about the choice of a criterion variable.

One stepwise multiple regression analysis was employed to reduce the initial battery of 12 predictor variables (predicting the average of Items 1-8 on the UCRSI) to an "optimally reduced subset" (see chapter I). The results of this regression analysis are presented in chapter IV and discussed in chapter V. As far as the procedure is concerned, this regression analysis provided a multiple correlation and an ordered, optimally reduced subset of predictors, both of which were needed as input to following steps.

A second stepwise multiple regression analysis was performed using the optimally reduced subset found in the first regression analysis and the five grading variables (see chapter I) to predict the same criterion (average of Items 1-8). First, the variables in the optimally reduced subset entered this regression equation in the same order that they entered the prior regression equation (i.e., ordered). Then, the grading variables were allowed to enter the regression equation in the order of their ability to improve the equation (i.e., floating). This served to test the incremental importance of the grading variables as predictors of ratings, given that certain other variables were already in the regression equation. Thus, the variables in the optimally reduced subset, in effect, "partialled out" a chunk of the variance before the grading variables were even considered.

The results of the second regression analysis also determined the relative importance of the grading variables and the overall ability of the available predictor variables to predict ratings. These results are also presented in chapter IV and discussed in chapter V.

Cross-validation of the multiple regression analyses was simulated through an examination of the estimated amount of shrinkage in the multiple correlations using McNemar's (1962, p. 184) shrinkage formula:

$$\underline{R}' = \{1 - (1 - \underline{R}^2)\ [(\underline{N} - 1) / (\underline{N} - \underline{n})]\}^{\frac{1}{2}} \qquad (2)$$

where $\underline{R}'$ is the multiple correlation after shrinkage, $\underline{N}$ is the sample size, $\underline{n}$ is the number of predictor variables, and $\underline{R}^2$ is the multiple correlation squared.

In order to test the significance of the increase in the multiple correlation when the grading variables were added to the optimal subset of other variables, an $\underline{F}$-test of significance was performed using another of McNemar's (1962, p. 284) formulae:

$$\underline{F} = \frac{(\underline{R}_1{}^2 - \underline{R}_2{}^2) \;/\; (\underline{m}_1 - \underline{m}_2)}{(1 - \underline{R}_1{}^2) \;/\; (\underline{N} - \underline{m}_1 - 1)} \qquad (3)$$

where $\underline{R}_1{}^2$ is the larger multiple correlation squared, $\underline{R}_2{}^2$ is the smaller multiple correlation squared, $\underline{N}$ is the sample size, $\underline{m}_1$ is the number of predictor variables associated with $\underline{R}_1$, and $\underline{m}_2$ is the number of predictor variables associated with $\underline{R}_2$, with degrees of freedom $\underline{m}_1 - \underline{m}_2$, and $\underline{N} - \underline{m}_1 - 1$.

The significance of the multiple correlations by themselves was determined with $\underline{F}$-tests of significance using yet another of McNemar's (1962, p. 283) formulae:

$$\underline{F} = (\underline{R}^2 \;/\; \underline{m}) \;/\; [(1 - \underline{R}^2) \;/\; (\underline{N} - \underline{m} - 1)] \qquad (4)$$

where $\underline{R}^2$ is the multiple correlation squared, $\underline{m}$ is the number of predictor variables included in the multiple correlation, and $\underline{N}$ is the sample size, with degrees of freedom $\underline{m}$ and $\underline{N} - \underline{m} - 1$.

The significance of the individual simple correlations was determined using a significance table for correlation coefficients.

Summary

Student ratings of college teachers at the University of Connecticut during the spring 1973 semester were studied to determine whether or not the addition of five new predictor variables dealing with grades could significantly improve an optimal set of predictors reduced from an initial battery of predictors. Factor analysis was used to reduce the eight-item rating instrument to a single criterion variable. Stepwise multiple regression analysis was used, both to reduce the initial battery of predictors to an optimally reduced subset, and to test the incremental importance of the grading variables as predictors of average ratings.

CHAPTER IV

RESULTS

This chapter presents the results of the statistical analyses of this study. The procedures used to obtain these results are explained in chapter III, and a discussion of these results is provided in chapter V.

Factor Analyses of the Rating Instrument's Items

Principal components factor analyses of two separate years' ratings (1972 and 1973) yielded nearly identical results. These results, presented in Table 5, show that all eight items on the University of Connecticut Rating Scale for Instruction (UCRSI) loaded heavily (.778 or greater) on a single factor. Item 8, "Over-All Summary As A Teacher," had a factor loading of over .97 both times. Furthermore, the correlations among the eight items, presented in Table 6, were all +.52 or greater and highly significant ($\underline{p} < .000001$). It should be noted that these correlations are among course section averages.

The results of these preliminary factor analyses indicated that there was essentially one, global dimension underlying the ratings data, and that a single criterion

Table 5

Factor Analyses of the Eight Items on the University of Connecticut Rating Scale for Instruction

	1972 Factor 1	1973 Factor 1
N	2417	2465
Eigenvalue	6.030	6.011
Percentage of Variance	75.372	75.139
Items:	Factor Loadings:	Factor Loadings:
1. Knowledge of Subject	.778	.783
2. Presentation of Material	.889	.895
3. Balance of Breadth & Detail	.887	.902
4. Enthusiasm for Subject	.852	.832
5. Fairness in Marking	.836	.824
6. Attitude Toward Student	.853	.852
7. Personal Mannerisms	.867	.863
8. Over-All Summary as a Teacher	.970	.971

Table 6

Product-moment Correlations Among the Eight Items
on the University of Connecticut
Rating Scale for Instruction for 1972 and 1973

Items:	1	2	3	4	5	6	7
1. Knowledge of Subject							
2. Presentation of Material	.66 (.66)						
3. Balance of Breadth & Detail	.69 (.66)	.88 (.87)					
4. Enthusiasm for Subject	.74 (.72)	.70 (.70)	.69 (.68)				
5. Fairness in Marking	.55 (.56)	.63 (.64)	.68 (.68)	.58 (.64)			
6. Attitude Toward Student	.52 (.54)	.66 (.65)	.66 (.64)	.66 (.70)	.81 (.81)		
7. Personal Mannerisms	.55 (.56)	.73 (.74)	.74 (.74)	.63 (.67)	.71 (.70)	.78 (.77)	
8. Over-All Summary as a Teacher	.74 (.74)	.90 (.89)	.89 (.87)	.79 (.81)	.77 (.77)	.81 (.81)	.82 (.83)

Note. 1972 correlations are in parentheses; N's for 1972 and 1973 were 2417 and 2465 respectively; all correlations are significant, $p \ll .000001$.

variable representing that factor could be employed in the subsequent stepwise multiple regression analyses. The average of all eight items was selected as the primary criterion for this study, but results were also obtained for three other criteria considered possibly representative of the factor (see below, Parallel Results Using Other Criteria). One of these other criteria, Item 8 by itself, was chosen on account of its extremely high factor loading. The average of the other seven items was chosen as a criterion for purposes of comparison with the first two criteria, since all eight items had very high factor loadings. Item 5 by itself was also used as a criterion in order to determine how well the predictor variables could predict the students' ratings of "grading fairness."

Reduction of the Initial Battery of Predictor Variables

The means and standard deviations of the 27 predictor and criterion variables are presented in Table 7. They were computed as part of this first stepwise multiple regression analysis. It should be noted that these means and standard deviations were computed across the 2,360 course sections and, therefore, that teachers are unequally represented according to the number of course sections they taught.

Table 8 shows the correlations among the 27 predictor and criterion variables. These are the correlations that were computed by the stepwise multiple regression analysis

Table 7

Means and Standard Deviations of the Predictor and Criterion Variables

	Variable	Mean	S.D.
1	Sex	1.15	.36
2	Appointment length	9.04	.28
3	Percentage employed	98.73	7.68
4	Class size	22.48	28.34
5	Number of years tenured	4.69	6.99
6	Age	41.31	10.10
7	Years since hiring	7.59	7.45
8	Course level	1.79	.79
9	Title level	2.65	1.00
10	Course location	1.80	.40
11	Department quantitativeness	33.75	43.35
12	Rate of return of ratings	.57	.16
13	Average grade	2.94	.57
14	Standard deviation of grades	.72	.32
15	Variance of grades	.61	.44
16	Skewness of grades	2.60	2.81
17	Kurtosis of grades	13.62	29.34
18	Item 1	8.12	1.13
19	Item 2	6.91	1.60
20	Item 3	6.87	1.47
21	Item 4	8.01	1.29
22	Item 5	7.61	1.29
23	Item 6	7.54	1.46
24	Item 7	7.52	1.38
25	Item 8	7.33	1.45
26	Average of Items 1-8	7.49	1.20
27	Average of Items 1-7	7.51	1.17

Note. N = 2,360 for each variable.

Table 8

Product-moment Correlations Among the Predictor and Criterion Variables

	Variable	1	2	3	4	5	6
1	Sex	1.00	-.03	-.07	.00	-.09	.08
2	Appointment length	-.03	1.00	-.10	.00	-.02	.00
3	Percentage employed	-.07	-.10	1.00	-.01	.07	.01
4	Class size	.00	.00	-.01	1.00	-.04	-.08
5	No. of years tenured	-.09	-.02	.07	-.04	1.00	.68
6	Age	.08	.00	.01	-.08	.68	1.00
7	Years since hiring	-.06	.00	.09	-.06	.95	.70
8	Course level	-.12	.00	-.01	-.14	.05	.08
9	Title level	-.20	-.03	.11	-.04	.62	.58
10	Course location	-.22	.06	-.01	.08	.15	-.03
11	Dept. quant.	-.09	-.02	.01	-.04	.08	.06
12	Rate of return	.05	-.01	-.02	-.06	.00	.02
13	Average grade	-.04	.08	-.01	-.16	.03	.03
14	Std. dev. of grades	.04	-.05	-.01	.20	-.03	-.07
15	Variance of grades	.04	-.05	.01	.14	-.03	-.07
16	Skewness of grades	.02	-.03	.01	.51	-.03	-.09
17	Kurtosis of grades	.02	-.02	.01	.77	-.03	-.07
18	Item 1	-.00	-.02	.08	-.06	.15	.16
19	Item 2	.05	.01	.05	-.02	.03	-.04
20	Item 3	.05	.00	.05	-.04	-.00	-.06
21	Item 4	.06	.01	.07	-.05	.05	.07
22	Item 5	-.02	.02	.01	-.06	-.03	-.04
23	Item 6	.02	.03	.03	-.07	.01	.00
24	Item 7	.06	.02	.05	-.06	-.03	-.07
25	Item 8	.02	.01	.06	-.05	.03	-.00
26	Average of Items 1-8	.04	.01	.06	-.06	.03	-.00
27	Average of Items 1-7	.04	.01	.06	-.06	.02	-.00

Note. All *r*'s ≥ .04 (or ≤ -.04) are significant, *p* < .05. For *r*'s ≥ .10 (or ≤ -.10), *p* < .000001.

Table 8 (Continued)

	Variable	7	8	9	10	11	12
1	Sex	-.06	-.12	-.20	-.22	-.09	.05
2	Appointment length	.00	.00	-.03	.06	-.02	-.01
3	Percentage employed	.09	-.01	.11	-.01	.01	-.02
4	Class size	-.06	-.14	-.04	.08	-.04	-.06
5	No. of years tenured	.95	.05	.62	.15	.08	.00
6	Age	.70	.08	.58	-.03	.06	.02
7	Years since hiring	1.00	.02	.61	.10	.07	.00
8	Course level	.02	1.00	.26	.34	-.01	.05
9	Title level	.61	.26	1.00	.30	.06	.01
10	Course location	.10	.34	.30	1.00	.04	-.04
11	Dept. quant.	.07	-.01	.06	.04	1.00	.06
12	Rate of return	.00	.05	.01	-.04	.06	1.00
13	Average grade	.02	.56	.15	.31	-.15	.03
14	Std. dev. of grades	-.03	-.52	-.15	-.17	.15	-.07
15	Variance of grades	-.03	-.47	-.14	-.17	.19	-.04
16	Skewness of grades	-.04	-.38	-.11	-.09	.18	-.04
17	Kurtosis of grades	-.03	-.21	-.05	-.01	.10	-.02
18	Item 1	.16	.11	.28	.02	-.01	.05
19	Item 2	.03	.15	.09	.06	-.09	.01
20	Item 3	.00	.16	.08	.04	-.03	.02
21	Item 4	.06	.14	.14	.01	-.12	.00
22	Item 5	-.03	.16	.06	.08	.02	-.01
23	Item 6	.01	.19	.05	.10	-.04	-.03
24	Item 7	-.03	.18	.06	.08	-.04	.01
25	Item 8	.04	.15	.12	.06	-.04	.01
26	Average of Items 1-8	.03	.18	.12	.07	-.05	.01
27	Average of Items 1-7	.03	.19	.12	.07	-.05	.01

Table 8 (Continued)

	Variable	13	14	15	16	17	18
1	Sex	-.04	.04	.04	.02	.02	-.00
2	Appointment length	.08	-.05	-.05	-.03	-.02	-.02
3	Percentage employed	-.01	-.01	.01	.01	.01	.08
4	Class size	-.16	.20	.14	.51	.77	-.06
5	No. of years tenured	.03	-.03	-.03	-.03	-.03	.15
6	Age	.03	-.07	-.07	-.09	-.07	.16
7	Years since hiring	.02	-.03	-.03	-.04	-.03	.16
8	Course level	.56	-.52	-.47	-.38	-.21	.11
9	Title level	.15	-.15	-.14	-.11	-.05	.28
10	Course location	.31	-.17	-.17	-.09	-.01	.02
11	Dept. quant.	-.15	.15	.19	.18	.10	-.01
12	Rate of return	.03	-.07	-.04	-.04	-.02	.05
13	Average grade	1.00	-.68	-.65	-.56	-.33	.14
14	Std. dev. of grades	-.68	1.00	.94	.78	.44	-.10
15	Variance of grades	-.65	.94	1.00	.86	.49	-.08
16	Skewness of grades	-.56	.78	.86	1.00	.82	-.07
17	Kurtosis of grades	-.33	.44	.49	.82	1.00	-.04
18	Item 1	.14	-.10	-.08	-.07	-.04	1.00
19	Item 2	.29	-.17	-.16	-.12	-.06	.67
20	Item 3	.27	-.16	-.15	-.11	-.06	.68
21	Item 4	.23	-.19	-.18	-.16	-.09	.73
22	Item 5	.41	-.20	-.17	-.14	-.10	.54
23	Item 6	.43	-.23	-.23	-.20	-.13	.51
24	Item 7	.32	-.18	-.17	-.14	-.08	.54
25	Item 8	.32	-.19	-.18	-.15	-.09	.74
26	Average of Items 1-8	.35	-.21	-.19	-.16	-.09	.77
27	Average of Items 1-7	.36	-.21	-.19	-.16	-.09	.77

Table 8 (Continued)

	Variable	19	20	21	22	23	24
1	Sex	.05	.05	.06	-.02	.02	.06
2	Appointment length	.01	.00	.01	.02	.03	.02
3	Percentage employed	.05	.05	.07	.01	.03	.05
4	Class size	-.02	-.04	-.05	-.06	-.07	-.06
5	No. of years tenured	.03	-.00	.05	-.03	.01	-.03
6	Age	-.04	-.06	.07	-.04	.00	-.07
7	Years since hiring	.03	.00	.06	-.03	.01	-.03
8	Course level	.15	.16	.14	.16	.19	.18
9	Title level	.09	.08	.14	.06	.05	.06
10	Course location	.06	.04	.01	.08	.10	.08
11	Dept. quant.	-.09	-.03	-.12	.02	-.04	-.04
12	Rate of return	.01	.02	.00	-.01	-.03	.01
13	Average grade	.29	.27	.23	.41	.43	.32
14	Std. dev. of grades	-.17	-.16	-.19	-.20	-.23	-.18
15	Variance of grades	-.16	-.15	-.18	-.17	-.23	-.17
16	Skewness of grades	-.12	-.11	-.16	-.14	-.20	-.14
17	Kurtosis of grades	-.06	-.06	-.09	-.10	-.13	-.08
18	Item 1	.67	.68	.73	.54	.51	.54
19	Item 2	1.00	.89	.70	.64	.66	.74
20	Item 3	.89	1.00	.69	.68	.67	.74
21	Item 4	.70	.69	1.00	.57	.65	.63
22	Item 5	.64	.68	.57	1.00	.81	.71
23	Item 6	.66	.67	.65	.81	1.00	.78
24	Item 7	.74	.74	.63	.71	.78	1.00
25	Item 8	.90	.89	.79	.78	.81	.83
26	Average of Items 1-8	.90	.91	.82	.83	.85	.87
27	Average of Items 1-7	.90	.90	.83	.83	.86	.87

Table 8 (Continued)

	Variable	25	26	27
1	Sex	.02	.04	.04
2	Appointment length	.01	.01	.01
3	Percentage employed	.06	.06	.06
4	Class size	-.05	-.06	-.06
5	No. of years tenured	.03	.03	.02
6	Age	-.00	-.00	-.00
7	Years since hiring	.04	.03	.03
8	Course level	.15	.18	.19
9	Title level	.12	.12	.12
10	Course location	.06	.07	.07
11	Dept. quant.	-.04	-.05	-.05
12	Rate of return	.01	.01	.01
13	Average grade	.32	.35	.36
14	Std. dev. of grades	-.19	-.21	-.21
15	Variance of grades	-.18	-.19	-.19
16	Skewness of grades	-.15	-.16	-.16
17	Kurtosis of grades	-.09	-.09	-.09
18	Item 1	.74	.77	.77
19	Item 2	.90	.90	.90
20	Item 3	.89	.91	.90
21	Item 4	.79	.82	.83
22	Item 5	.78	.83	.83
23	Item 6	.81	.85	.86
24	Item 7	.83	.87	.87
25	Item 8	1.00	.97	.96
26	Average of Items 1-8	.97	1.00	1.00
27	Average of Items 1-7	.96	1.00	1.00

subprogram of the Statistical Package for the Social Sciences (SPSS, Nie et al., 1970). These correlations were subsequently used by that subprogram to perform the regression analyses. It should be noted that these are, again, correlations among course section averages. Also, because more cases were deleted for missing data for the regression analyses (more variables involved) than for the factor analyses, some of the correlations in Table 6 differ very slightly from the corresponding ones in Table 8.

Given the large sample size, any of these correlations that exceeds .035 is significantly different from zero ($p < .05$). Furthermore, for any r that exceeds .048, p is less than .01; and if r exceeds .10, then p is less than .000001. Thus one may be fairly certain that even the relatively weak relationships found in this study were not attributable to chance variation.

The first major finding, bearing on the research question, was the correlation between the average student grade in each course section and the average student rating of the teacher of that course section. This was found to be .35 ($p \ll .000001$). The other correlations involving these two variables are particularly interesting. For example, a correlation of -.15 was found between the average grade in each course section and the quantitativeness of the department in which that course is taught. Also, the correlation between average grade and Item 5 on the rating instrument ("Fairness in Marking") was .41, and the correlation between average grade and Item 6 ("Attitude Toward Student") was .43. Discussion of the implications of these and other results is withheld until chapter V.

The results of the first stepwise multiple regression analysis are presented in Table 9. The initial battery of 12 predictors of ratings was reduced to an "optimally reduced subset" (see chapters I and III). This subset consisted of the first 10 variables shown in Table 9 (above the line of dashes). The variables are listed in the rank order of their importance in improving the predictability of ratings by the regression equation. Also shown in Table 9, for each variable, are the standard error of estimate after the variable's inclusion in the regression equation, the multiple $\underline{R}$, $\underline{R}^2$, the increase in $\underline{R}^2$ over the previous step, the simple correlation with the criterion, and the $\underline{F}$-value that signifies the importance of the variable to the regression equation as of the last step.

The multiple correlation produced by the optimally reduced subset of 10 predictors was .25. Correction for shrinkage yielded an $\underline{R}$ of .24. Although this multiple correlation is not very large (and accounts for only slightly more than 6% of the criterion variance), it is highly significant, $\underline{F}$ (10, 2349) = 14.12, $\underline{p} < .001$. It should be noted that the first variable to enter the regression equation, "course level," accounted for over half of the criterion variance finally accounted for by the entire optimally reduced subset of 10 predictors.

Table 9 First Stepwise Multiple Regression Analysis—Reduction of the Initial Battery of Predictor Variables

Rank	Variable	R	R^2	Increase in R^2	Standard Error of Estimate	Simple r	F-Value	Signif.
1	Course level	.18162	.03299	.03299	1.18363	.18162	52.190	***
2	Title level	.19699	.03881	.00582	1.18032	.12155	31.695	***
3	Age	.21045	.04429	.00548	1.17719	-.00118	19.545	***
4	Sex	.22995	.05288	.00859	1.17214	.03509	19.702	***
5	Percentage employed	.23486	.05516	.00228	1.17098	.05653	5.181	*
6	Dept. quant.	.23889	.05707	.00191	1.17004	-.05062	4.775	*
7	Class size	.24261	.05886	.00179	1.16918	-.05908	3.772	*
8	Appointment length	.24366	.05937	.00051	1.16911	.01186	1.120	n.s.
9	Years since hiring	.24431	.05969	.00032	1.16916	.03132	2.065	n.s.
10	No. of years tenured	.24555	.06029	.00061	1.16903	.02543	1.242	n.s.
11	Course location	.24603	.06053	.00024	1.16913	.06555	.605	n.s.
12	Rate of return	.24605	.06054	.00001	1.16938	.00768	.014	n.s.

*** $p < .001$ * $p < .05$ n.s. $p > .05$

Furthermore, even though all 12 predictor variables increased the multiple R somewhat, only the first 10 reduced the standard error of estimate (optimality).

The F-values of the predictors as of the last step show the final significance of each predictor. They also demonstrate the fact that predictors may gain or lose significance when other predictors enter the regression equation. (The rank order of the F-values is not the same as the order of inclusion of the variables into the regression equation.)

Addition of the Five New Predictor Variables to the Optimally Reduced Subset

The results of the second stepwise multiple regression analysis are presented in Table 10. These results show that the average of the student grades in each course section was the single best predictor of the average rating of the teacher of that course section. Furthermore, the addition of the grading variables to the optimally reduced subset of other predictors significantly improved the multiple correlation from .25 to .39, $F\ (4, 2345) = 60.13$, $p < .001$. The variable "average grade" by itself accounted for nearly 8.5% of the criterion variance (more than was accounted for by the entire optimally reduced subset of other predictors).

A further indication of the importance of the grading variables is provided by the fact that the variable "average grade," when it entered the regression equation at step

Table 10 Second Stepwise Multiple Regression Analysis—Addition of the Five Grading Variables to the Optimally Reduced Subset

Rank	Variable	R	R^2	Increase in R^2	Standard Error of Estimate	Simple r	F-Value	Signif.
1	Course level	.18162	.03299	.03299	1.18363	.18162	1.333	n.s.
2	Title level	.19699	.03881	.00582	1.18032	.12155	29.842	***
3	Age	.21045	.04429	.00548	1.17719	-.00118	11.949	***
4	Sex	.22995	.05288	.00859	1.17214	.03509	17.323	***
5	Percentage employed	.23486	.05516	.00228	1.17098	.05653	6.003	**
6	Dept. quant.	.23889	.05707	.00191	1.17004	-.05062	.002	n.s.
7	Class size	.24261	.05886	.00179	1.16918	-.05908	3.878	*
8	Appointment length	.24366	.05937	.00051	1.16911	.01186	.177	n.s.
9	Years since hiring	.24431	.05969	.00032	1.16916	.03132	3.167	*
10	No. of years tenured	.24555	.06029	.00061	1.16903	.02543	3.298	*
11	Average grade	.38080	.14501	.08472	1.11533	.35333	206.038	***
12	Skewness of grades	.38445	.14780	.00279	1.11375	-.15866	3.191	*
13	Standard deviation of grades	.38469	.14799	.00018	1.11386	-.20510	3.532	*
14	Variance of grades	.38621	.14916	.00117	1.11334	-.19352	3.168	*
15	Kurtosis of grades	.38622	.14916	.00001	1.11357	-.09191	.019	n.s.

*** $p < .001$ ** $p < .01$ * $p < .05$ n.s. $p > .05$

number 11, completely dominated the regression equation. It took over and rearranged the regression equation to such an extent that the variables "course level" and "department quantitativeness" lost nearly all of their significance as contributors to the final regression equation. The extent of the rearrangement is clearly indicated by the column of F-values in Table 10 compared to the same column in Table 9.

The first 14 variables listed in Table 10 (above the lower line of dashes) produced the multiple correlation with the lowest standard error of estimate. That R was .39. Correction for shrinkage yielded an R of .38. While this is still not a very large multiple correlation (accounting for only about 15% of the criterion variance), it is highly significant by itself, F (14, 2345) = 28.28, $p < .001$. The significance of the increase in the multiple correlation is described above.

Parallel Results Using Other Criteria

The choice of the criterion variable for the above regression analyses is explained above (see Factor Analyses of the Rating Instrument's Items). As mentioned above, three other criteria, Item 8 by itself, the average of the other seven items, and Item 5 by itself, might represent the single ratings factor as well as the average of all eight items. In order to determine if the choice of the criterion was important, three more pairs of stepwise multiple regression analyses were performed. These analyses were run for the three other criteria exactly as they were for the first criterion. A reduction of the initial battery of predictors was done first, and then the five grading variables were added to the new optimally reduced subset in each case.

The results of these analyses were nearly identical to those obtained with the original criterion. Using "Item 8" as the criterion, the reduction of the initial battery of predictors resulted in a multiple correlation of .23. The correction for shrinkage yielded an R of .22, F (9, 2350) = 13.01, $p < .001$. The variable "appointment length" did not make a significant contribution to this regression equation, though it did for the other three criteria. Thus, this optimally reduced subset of predictors of "Item 8" included only nine of the independent variables.

The addition of the grading variables to this optimally reduced subset of nine predictors increased the multiple correlation to .36, F (13, 2346) = 26.09, $p < .001$. The correction for shrinkage did not lower this R. The _increase_ in the multiple correlation was also highly significant, F (4, 2346) = 52.94, $p < .001$.

Similarly, when the average of the first seven items on the UCRSI was used as the criterion, the reduction of the initial battery of predictors resulted in a multiple correlation of .25. Correction for shrinkage yielded an R of .24, F (10, 2349) = 14.44, $p < .001$. This optimally reduced subset of predictors included the same 10 variables as the subset of predictors of the average of all eight items, and there was only one minor difference in their order: the variables "age" and "sex" were reversed although their respective F-values as of the last step were nearly identical in both analyses.

The addition of the five grading variables in this case increased the multiple correlation to .39, with the correction for shrinkage yielding an R of .38, F (14, 2345) = 28.72, $p < .001$. The increase in the multiple correlation was also highly significant, F (4, 2345) = 60.75, $p < .001$.

The use of "Item 5" as the criterion yielded slightly different results. The reduction of the initial battery of predictors produced a lower multiple R than it did for the other three criteria. However, it was still a highly significant .19, F (9, 2350) = 9.29, $p < .001$. The correction for shrinkage did not lower this R. The resulting optimally reduced subset consisted of nine of the predictors. "Course level," "age," and "title level" were the best three predictors out of the initial battery of 12 (as in the other reductions), but the order of the less important predictors was altered.

When the five grading variables were added to the optimally reduced subset of predictors of "Item 5," the multiple correlation increased to .45, F (13, 2346) = 44.87, $p << .001$. The correction for shrinkage did not lower this R. This increase (from the smallest R found in this study to the largest) was, of course, very significant,

F (4, 2346) = 120.71, p << .001.

It should be noted that for all four pairs of regression analyses, the reported F-values computed for the increases in the multiple R's (on account of the addition of the grading variables) were based on the multiple R's after shrinkage. This is conservative (slightly understating the F-value), but in this study it resulted in no practical differences, since all of the F-values were so highly significant.

Summary

Factor analyses of the eight-item rating instrument showed that there was essentially one factor underlying the UCRSI ratings data. This led to the choice of the average of all eight items on the UCRSI as the criterion variable to represent that factor. Multiple regression analyses yielded low but highly significant multiple correlations. Moreover, they showed that the addition of the grading variables to the optimally reduced subset of other predictors of ratings did indeed make a highly significant improvement in the predictive efficiency of the regression equation. Furthermore, practically identical results were obtained for three other criterion variables.

CHAPTER V

DISCUSSION AND CONCLUSIONS

This chapter presents a discussion of the results of this study and explores several implications of those results. This chapter also provides some recommendations in light of certain conclusions based on those results and implications.

Discussion of Results and Implications

The results of this study apparently provide a unique contribution to the literature on the research question: What is the influence of the grades students receive on their ratings of the college teachers who gave them those grades? This study has used multivariate techniques, and has studied a very large sample of course sections across an entire university. Previous studies have studied typically only one or a few course sections, or have suffered from several other inadequacies (see chapter I). This study did not suffer from those inadequacies, and thus the results are probably more definitive and generalizable. Furthermore, these results provide more up-to-date knowledge in view of certain changes that have occurred in rating and grading practices over the past decade (see chapter I).

The primary implication of the results of this study is that there is an interrelationship between grades and ratings. It would seem that both the direction and the extent of the hypothesized "grading bias" have been demonstrated. That is, students apparently tend to rate lower those teachers from whom they receive lower grades and vice versa. However, in light of the fact that the grading variables accounted for only about 9% of the variance in ratings in this study, other factors (valid or not) must also be influencing the students' ratings. In other words, students did not simply "payoff" their teachers with ratings in direct proportion to the grades they received from those teachers. Rather, the students were biased or influenced by their grades, overall, in such a way that permitted other considerations also to be involved in the rating process.

Possibly some of the students did strictly "payoff" their teachers for grades received, while others were not at all influenced by their grades. On the other hand, it is possible that most or all of the students' ratings are merely "shifted" by the grades received. That is, students might rate teachers more or less validly, but plus or minus a certain amount according to the grades received. Without the ability to pair individual student grades and ratings (given the anonymity of ratings), it could not be determined whether some students were significantly more influenced by their grades than were other students. This would be an appropriate question for further research to answer conclusively, but the previous studies that have used paired data have often found essentially the same degree of relationship between grades and ratings, in spite of certain inadequacies or faults in those studies.

Overall average ratings are often the measure used by administrators and department chairmen for making decisions about faculty pay, promotion, and tenure, and this study has shown the probable existence of an overall grading bias. To some extent, it is irrelevant what the differences are between individual students, when it comes to the grading bias. It does not even matter whether the bias is conscious or subconscious, so long as it is actually dependent upon grades. Unless one could develop a way to count only the ratings of those students who are not biased by grades, or unless a conscious grading bias were subject to elimination more than an unconscious bias, then the overall grading bias will exist whatever its makeup might be.

Even though the rate of return of the ratings used in this study was only 55%, it has been about 55% for many years, and these are the ratings that are systematically used for making administrative decisions about faculty careers. That is, even though returned ratings may not be generalizable across all of the students the faculty member taught (since students who return ratings may not constitute a random sample so far as their opinions are concerned), such ratings are commonly used as though they were representative of overall student opinion. The results of this study should be generalizable to the many other rating systems operating with similar rates of return. It remains for future researchers to determine whether an appreciably different rate of return would have any influence on the relationships found in this study, but these results are applicable to rating systems *as they are commonly used*.

Also, to the extent that the course sections and rating procedures at the University of Connecticut are representative of those across the nation, the results of this study are generalizable. The relationship between grades and ratings probably varies from one institution to another because of differences in the students, faculty, grading systems, rating instruments, and rating procedures. For example, the timing of ratings (before or after final grades are awarded) is an important difference between rating systems, even though previous research (e.g., Bausell & Magoon, 1972a; Holmes, 1971) indicates that an *expected grade* bias probably exists before the actual grade is determined. Perhaps future researchers could use similar rating procedures at a large number of institutions with comparable grading practices to find out how generalizable the results of this study are.

It is possible that there are other variables which, when added to the final regression equation in this study, would lower or eliminate the importance of the grading variables as predictors of ratings. That is, it is possible that grades and ratings correlate without there being any causal relationship between them, and that ratings actually depend completely on some other unknown factor or factors. This is doubtful, however. The simple and multiple correlations found in this study indicate a definite, if partial, relationship. Furthermore, the very high significance levels of these simple and multiple correlations suggest that the results are not attributable to chance variation. Also, certain previous findings (see especially Holmes, 1972) provide strong (experimental) evidence of a causal relationship between grades and ratings.

The results of this study would have been even more definitive, had the optimally reduced subset of the initial battery of predictors accounted for a larger amount of the variance in ratings. This would have lowered the chances that any other variable exists which would account for most of the variance in the ratings (thus suggesting that the present results might be spurious). The more criterion variance "partialled out" by the optimally reduced subset, the more significant the increase attributable to the grading variables would have been.

Ideally, nearly all of the variance in ratings would be attributable to students' reliable perceptions of their teachers' performances (Treffinger and Feldhusen, 1970).

If the grading bias were about the only exception to this rule, it might explain why only about 6% of the criterion variance could be accounted for by the 10 predictors in the optimally reduced subset. That is, maybe there are no predictors of ratings more significant than those found in this study (outside of other variables which might tap the students' opinions about their teachers in some other way).

Even with the grading variables added in, the final regression equation in this study was able to account for only about 15% of the criterion variance. With such a large portion of the variance left unaccounted for, the chances are theoretically greater that some variable could disprove the existence of what seems to be the grading bias. This writer doubts the existence of any such variable. Most likely there are few accurate predictors of student ratings of faculty performance.

One possibility is that student ratings are almost totally invalid as measures of effective teaching anyway, and relate more to student and faculty personality interactions. Similarly, peer ratings may be measures of personality conflicts or even secondhand student ratings. Furthermore, without a consensual definition of effective teaching (or the purposes of education for that matter), even unbiased judgments may not be highly related to each other.

The results of studies by Treffinger and Feldhusen (1970) and by Bausell and Magoon (1972b) indicate that students' first impressions and preconceptions are persistent and may be the best predictors of end-of-course ratings. Thus, it is possible that researchers should be looking for predictors of such first impressions or preconceptions. Perhaps the grades students receive constitute the major influence on ratings after the initial opinion is formed.

Also, to the extent that there is a large portion of error variance in the ratings data, the importance of the grading bias found in this study looms even larger. That is, if the reliability of the ratings is considerably lower than 1.00, then the grading variables would account for more than 9% of the systematic variance. For example, if the reliability were .75, then the percentage of the systematic variance accounted for by the grading variables would be 12% (.09 / .75). Unfortunately, there is no accurate estimate of exactly how reliable student ratings are, nor even how reliability would best be defined. But for the purposes of this study, it is conservative to assume that the reliability is nearly 1.00 and that the grading bias is no more significant than indicated by the findings.

Several of the details of the results of this study are worthy of mention. Notably, the results of the factor analyses of the eight-item rating instrument (see Tables 5 and 6) are interesting in their own right, beyond their utility in the selection of criteria. Whatever the global impression of teachers may depend upon, it was apparently measured in the same way by the rating instrument in 1972 and 1973. The single factors and the correlations between items for the two years are strikingly similar. It is also interesting to note the order of the sizes of the factor loadings. "Knowledge of Subject" had the lowest factor loading of all eight, and "Over-All Summary as a Teacher" had the highest loading.

The finding of only one factor suggests that a halo effect is present, and, as suggested by Hoyt (1969) and by Widlak et al. (1973), the separate items may be of little diagnostic value. This writer doubts that the high correlations found between items represent any "true" relationships between the eight traits that the instrument attempts to measure. Rather, there seems to be a halo effect confounding the meaning of the separate items. It is possible that certain "profile effects" (indicative of differences across the rating items) may exist, even though masked by the halo effect. However, the diagnostic value of such "high inference" items (very general and open to varying interpretations) is questionable anyway, even if there were no confounding influences. It is not at all clear exactly what a teacher should do to improve his ratings on such global traits.

Student ratings, therefore, may not be improving teaching in the expected way. Specific, but global, items apparently can not help the teacher improve his teaching. In fact, with a grading bias present, ratings almost assuredly act counter to the larger educational goals. That is, while a full range of grades may be educationally appropriate, the grading bias would act to diminish the apparent effectiveness of teachers who do give out some low grades, and it would act to reward the lenient teachers who give out higher and less discriminating grades.

Of course, one optimistic explanation for a grading bias is that many high grades accompanied by high ratings in a course section might reflect superior teaching or a highly successful class (perhaps using "contract grading"). However, this writer doubts that such is the general case. External criteria, such as standardized achievement tests, would be needed to substantiate any claim that such high grades are deserved, especially in light of the grading trend over the past several years (see chapter I).

It was hypothesized that certain departments at this university were more lenient (in the awarding of grades) than others. Specifically, it was thought that the "hard" science and mathematical departments were giving out lower grades than other departments, and that the faculty members in those highly quantitative departments might be suffering from lower ratings. The correlation of -.15 found between average grade and department quantitativeness supports such a theory, as does the correlation of -.05 between ratings and department quantitativeness. Of course, these correlations do not prove any causality of the relationships; in addition, the size of the correlations suggests that the relationships are slight.

Nevertheless, department quantitativeness was a significant predictor of ratings until the grading variables were allowed to enter the regression equation; then its predictive potential was subsumed by the grading variables. It would seem that the variable "department quantitativeness" was providing some information about grades indirectly (because of the relationship between department quantitativeness and grades). However, when "better" information about grades (the grading variables themselves) entered the regression equation, then department quantitativeness was of no further importance in predicting ratings. The same situation occurred with the variable "course level," which was the best predictor of ratings before the grading variables were considered. The *F*-values in Tables 9 and 10 demonstrate the extent of these variables' loss in predictive potential.

It is possible that certain departments are suffering more from the grading bias than others on account of the differences in grading practices. Moreover, it is likely that some individual teachers are suffering more than others according to their grade distributions. The teachers with the most lenient grade distributions are probably not always the best teachers. Thus, the grading bias lowers the probability that ratings could be valid as measures of teaching effectiveness.

As suggested earlier, the effect of grades on ratings well may be only partial, i.e., one of several influences (see discussion of first impressions above). The simple correlation between grades and ratings found in this study was not high (.35), but it was very highly significant because of the large sample size. Many previous studies, in spite of serious shortcomings (see chapter I), also found significant correlations of this approximate magnitude. Typically the correlations did not exceed .30 (Costin et al., 1971). The results of this study support the findings of the large portion of the previous research which found such correlations. Thus, this correlation of .35 between grades and ratings, and the similar correlations found by other researchers, could be quite accurate and indicate that the "true" relationship between grades and ratings probably lies in the vicinity of .30 to .35.

The correlation of .41 found between average grade and Item 5 ("Fairness in Marking") indicates that this item is especially sensitive to the grading bias. Students evidently consider higher grades as "fair." This is not very surprising. One might expect that a rating item like "Fairness in Marking" would receive the brunt of the grading bias. Actually, however, the grading bias apparently affects all eight items (see correlations among average grade and Items 1-8 in Table 8). Furthermore, the correlation is highest (.43) for Item 6 ("Attitude Toward Student"). Students apparently consider grades awarded as a primary indication of their teachers' attitudes toward students.

While Items 5 and 6 seem to be the most sensitive of the eight items to the grading bias (having the highest correlations with all five of the grading variables), other considerations must also enter into the students' ratings even of these two items. When Item 5 by itself was the criterion variable (see Parallel Results Using Other Criteria), the grading variables still managed to account for only about 17% of the criterion variance. Thus, the bias is partial even for Items 5 and 6, but it is pervasive across all eight items. This finding supports the above mentioned assertion by Bausell and Magoon (1972b) that disappointed students "will tend to deprecate the instructor's teaching performance in areas other than his grading system" (p. 130, emphasis added).

Conclusions and Recommendations

It would seem that one important influence on student ratings of college teachers has been identified. It is the so-called "grading bias," and it apparently accounts for about 9% of the variance in ratings. Variables which could account for most of the rest of the variance, however, have not been identified. Student ratings may or may not be mostly valid in spite of the grading bias. There is no proof either way. It remains for future research to answer many such questions raised by the results of this study. Even if researchers could find variables which would account for the rest of the variance in the ratings, that would be no guarantee that ratings are valid as measures of effective teaching. Such variables, however, probably would offer substantial clues as to whether or not ratings are valid, depending on the identity of those variables.

Well defined educational goals are needed before effective teaching can be defined, and external criteria of effective teaching must be well defined before valid measures of effectiveness can be firmly established. This positive, constructive, and difficult work needs to be done.

The results of this study are, therefore, somewhat negative. That is, a grading bias has been found which most likely serves to lower the validity of student ratings. Perhaps positive steps could be taken which would eliminate the grading bias. If possible, methods should be devised which would do just that. One possibility is that ratings could be collected very early in the semester. This would allow for feedback to the teachers in time for improvement to occur before the end of the semester (assuming the teachers would be responsive and that ratings would indicate desired changes). However, there is still likely to be an expected grade bias (Holmes, 1971), and there is still no certainty about the validity of ratings even without any grading bias.

The results of this study have demonstrated a grading bias for a "high inference" rating instrument (containing very general questions open to varying interpretations), but radically different results might occur with so-called "low inference scales" (asking more objective questions which supposedly would be less subject to halo effects and biases). Such scales may or may not eliminate the grading bias and may or may not be valid as measures of effective teaching. These facts remain to be determined by future researchers.

One might suggest that students could be "taught" how to rate their teachers more fairly. The evidence of a halo effect, however, suggests that students are unable to separate their feelings about grades and teachers, especially on "high inference scales." Or it might be suggested that, if grading practices were completely fair and non-arbitrary, then any bias caused by grades would disappear. However, if grades were made fair (i.e., more justly discriminating), it seems that grades would be lower, and that this might in turn increase the bias.

Perhaps the most important conclusion for immediate consumption is that, since grades do seem to influence student ratings of college teachers, this bias should be taken into account whenever one is interpreting student ratings. Administrators, department chairmen, and promotion and tenure committees across the nation should remember that the grading bias exists and that certain teachers may suffer from the bias more than others, depending on the grades they give out. It seems intuitively obvious that one should not reward teachers for leniency if students are expected to work and to achieve. It would be better to reward teachers according to the actual achievement levels reached by their students (measured by standardized achievement tests).

One policy option might be to drop student ratings altogether. They are costly, and possibly not worth the cost, especially since they must be interpreted with caution. Some reasons for not eliminating student ratings are: they are well established; some institutional prestige is associated with their use; many students want them; they at least give the appearance of requiring teacher accountability; and dropping them could conceivably lower faculty concern for effective teaching. On the other hand, it might be argued, better teaching and fairer grading would occur without the faculty's probable fear of reprisals on ratings (Ladas, 1974; "Too Many A's," 1974).

This writer recommends that university officials and faculties review the above implications, and decide how best to serve the educational needs involved. If student ratings can not be made objective and valid, if the grading bias can not be eliminated, and if student ratings can not be dropped outright, then it would seem two possible paths are suggested. Either decision makers should ignore the ratings, or they should combine them with other measures of teacher effectiveness (perhaps achievement tests or indices of the amount of work done by the students) in such a way that ratings would not encourage teachers to be slack or to demand too little.

APPENDIX

MACRO FLOWCHARTS

OF THE DATA PROCESSING STEPS USED IN THIS STUDY

Figures 6 through 18 represent the major data processing steps required to accumulate the data for this study. Figure 5 provides a key to the symbols used in those figures, and the letters and numerals within the symbols in the figures are the names of the tapes, documents, and operations symbolized. Keypunching and manual data lookup operations are labeled but not named. A brief description of the purpose of each step follows.

Merge 1. This first step matched and merged data from the "instructor header record" tape (one instructor header record per course section evaluated) with data from the "professional history file" tape. The merged data were output onto both paper and tape, and information about missing data was also printed on the paper output for use in the next step.

First missing data input. In this step, the information obtained in the merge 1 step was used to look up and punch onto cards the data missing from the professional history file tape. The resulting deck of cards was retained for input into the merge 5 step below.

Merge 2. This step matched and merged the faculty evaluations data with the merged output from the merge 1 step. The individual ratings data were used to compute means, N's, and standard deviations for each item on the rating scale, and the results were output onto both paper and tape.

Sort 1. The records on the output tape from the merge 2 step above were sorted by branch, department, course number, and section number. The sorted records were output onto another tape for later use in the merge 3 step.

Sort 2. The grade distribution data was sorted as in sort 1 above so that the course sections would be in the same sequence on both tapes for input into the merge 3 program.

Merge 3. This program matched and merged the grade distribution data with the other data previously assembled for each course section. Computations of average grades and standard deviations of grades in the course sections were made at this point, and the merged results were output onto paper and tape. Also on the paper output was information about missing data detected by this program.

Second missing data input. The information on missing data from merge 3 was used, in this step, to create a deck of cards containing that missing data. The cards were retained for input into the merge 5 step below.

Merge 4. This program matched students' records of their Quantitative and Verbal Scholastic Aptitude Test scores with their declared majors in order to compute for each department the variable "department quantitativeness." The results of these computations were output onto paper and cards. The deck of cards was retained for input into the merge 5 program.

Merge 5. The purpose of this step was to insert the data on cards (from previous steps) into the course section records, to compute several new variables from old ones (age from date of birth for example), and to edit the data for the acceptability of the values. The results were output onto paper and tape, and information about missing data and improper values was also printed on the output paper.

Third missing data input. The information on missing or unacceptable data from the merge 5 program was used to create a deck of cards containing the missing data. This deck was retained for input into the merge 6 program below.

Merge 6. This program was used to insert the missing data discovered in the merge 5 step into the final data records, which were output onto paper and tape.

Regression run 1. This step accomplished the reduction of the initial battery of predictor variables using the stepwise multiple regression analysis subprogram of the Statistical Package for the Social Sciences (SPSS, Nie et al., 1970). The results of this analysis were used to make the ordered list of variables (the reduced set in the order of their inclusion in the regression equation) for input into the next step.

Regression run 2. This step was another stepwise multiple regression analysis using SPSS as above, but with the five grading variables added to the optimally reduced subset of predictors found in regression run 1.

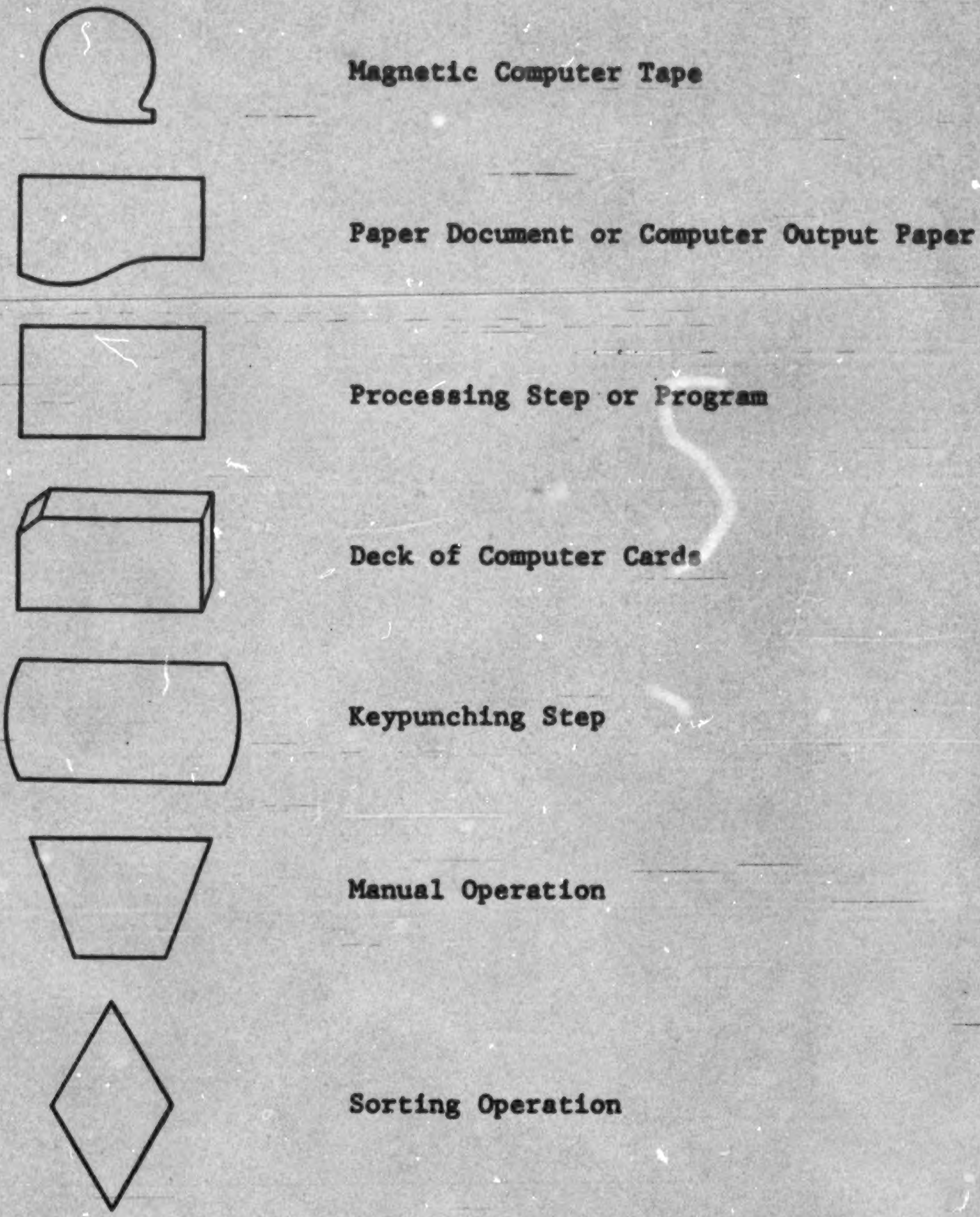
Magnetic Computer Tape
Paper Document or Computer Output Paper
Processing Step or Program
Deck of Computer Cards
Keypunching Step
Manual Operation
Sorting Operation

Figure 5. Key to symbols used in Figures 6 through 18.

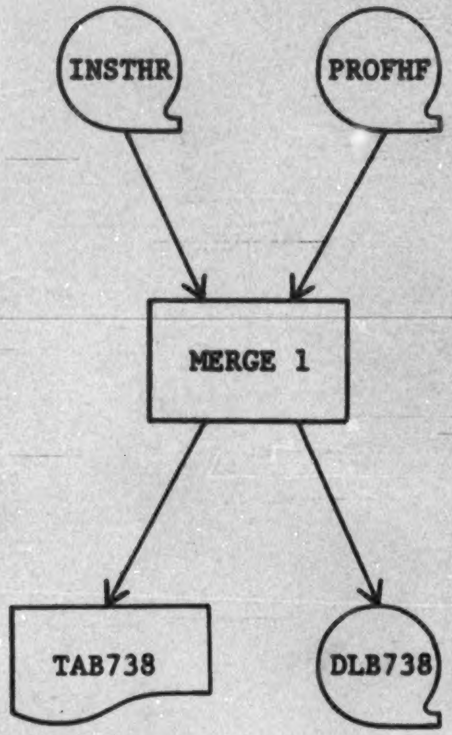
INSTHR
PROFHF
MERGE 1
TAB738
DLB738

Figure 6. Merge 1.

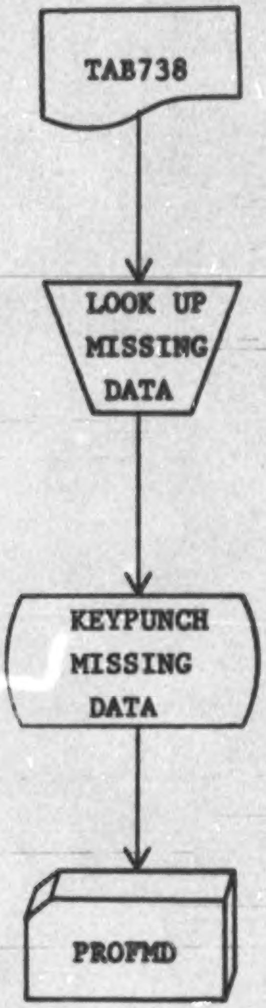
TAB738
LOOK UP MISSING DATA
KEYPUNCH MISSING DATA
PROFMD

Figure 7. First missing data input.

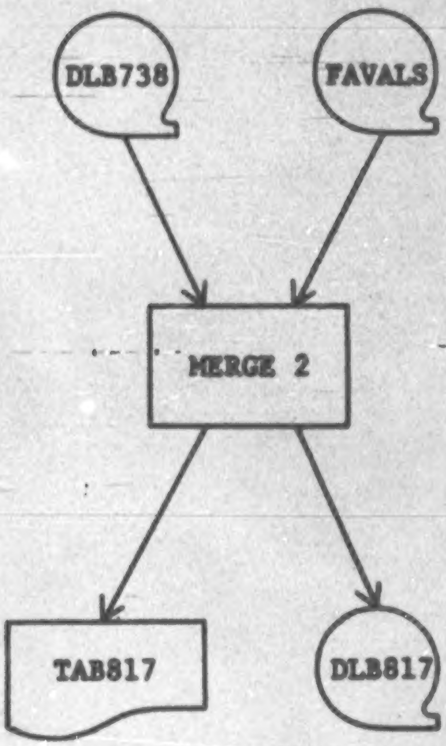
DLB738
FAVALS
MERGE 2
TAB817
DLB817

Figure 8. Merge 2.

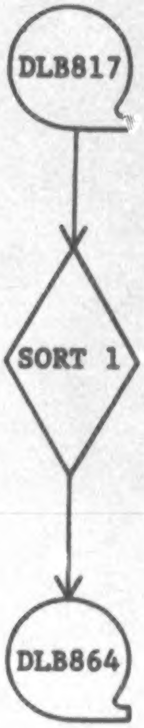
DLB817
SORT 1
DLB864

Figure 9. Sort 1.

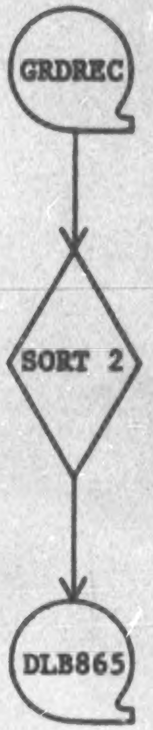
GRDREC
SORT 2
DLB865

Figure 10. Sort 2.

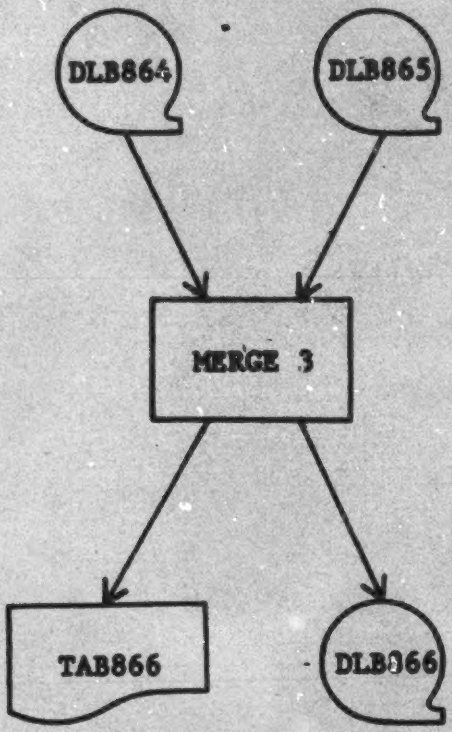
DLB864
DLB865
MERGE 3
TAB866
DLB866

Figure 11. Merge 3.

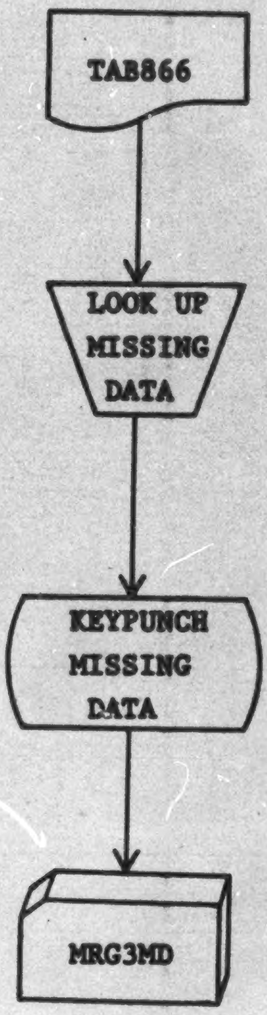
TAB866
LOOK UP MISSING DATA
KEYPUNCH MISSING DATA
MRG3MD

Figure 12. Second missing data input.

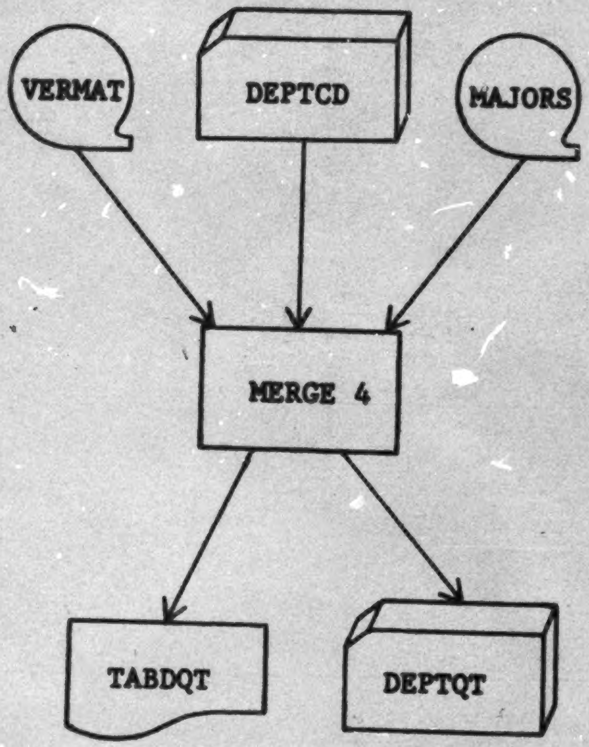
VERMAT
DEPTCD
MAJORS
MERGE 4
TABDQT
DEPTQT

Figure 13. Merge 4.

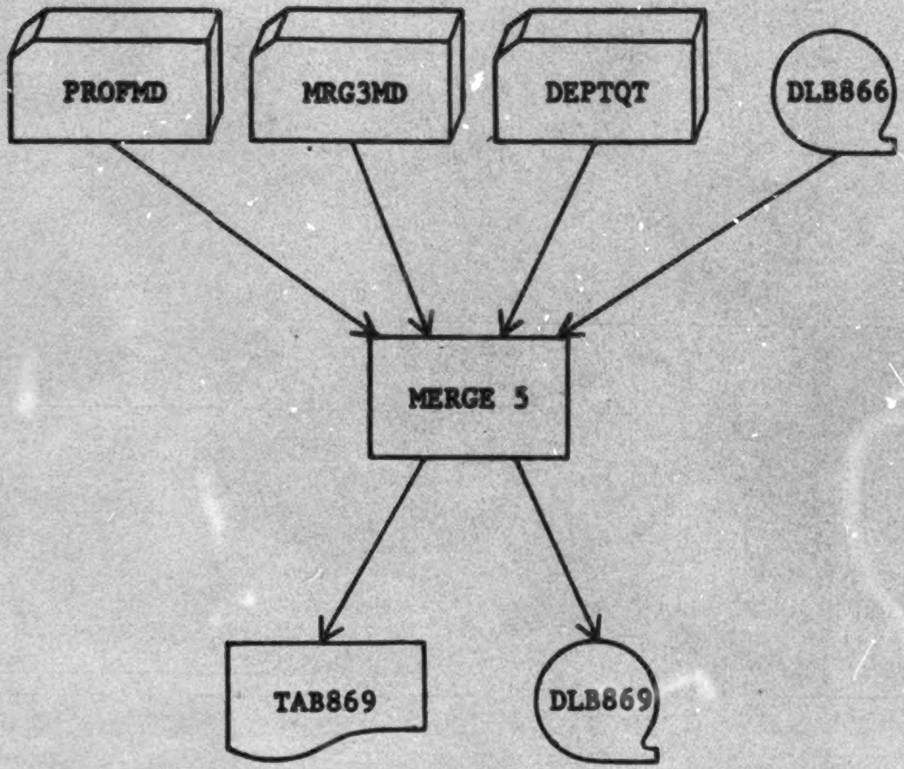
PROFMD
MRG3MD
DEPTQT
DLB866
MERGE 5
TAB869
DLB869

Figure 14. Merge 5.

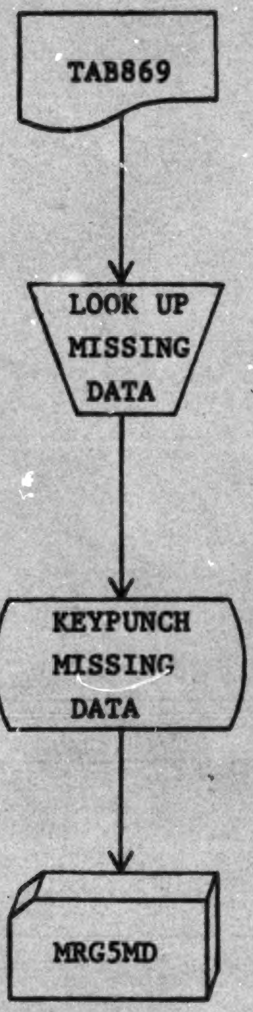
TAB869
LOOK UP MISSING DATA
KEYPUNCH MISSING DATA
MRG5MD

Figure 15. Third missing data input.

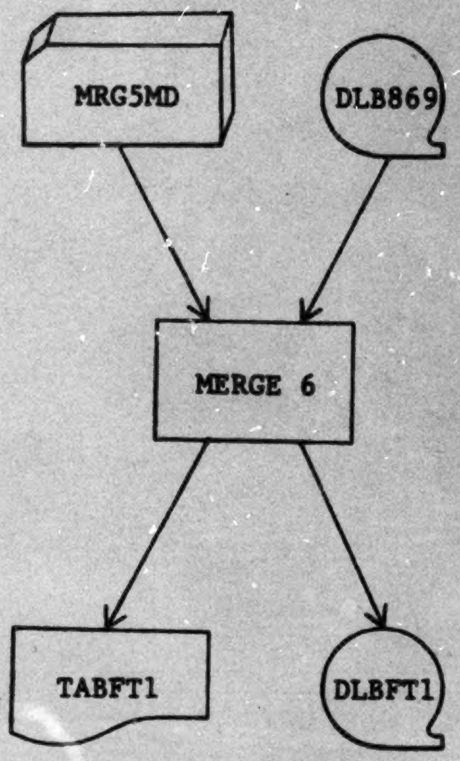
MRG5MD
DLB869
MERGE 6
TABFT1
DLBFT1

Figure 16. Merge 6.

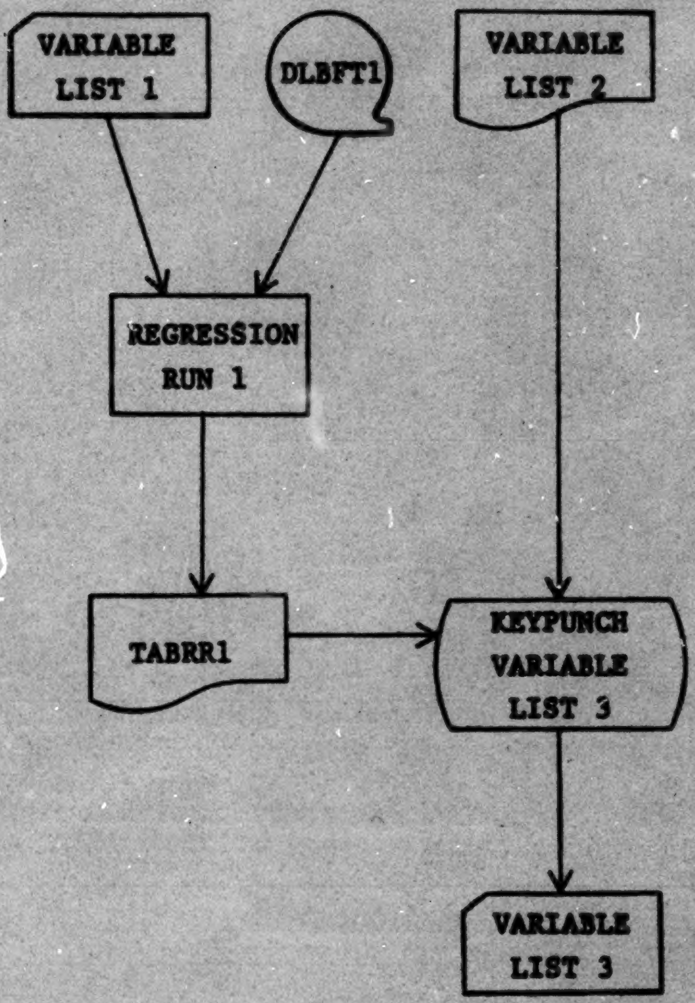
VARIABLE LIST 1
DLBFT1
VARIABLE LIST 2
REGRESSION RUN 1
TABRR1
KEYPUNCH VARIABLE LIST 3
VARIABLE LIST 3

Figure 17. Regression run 1.

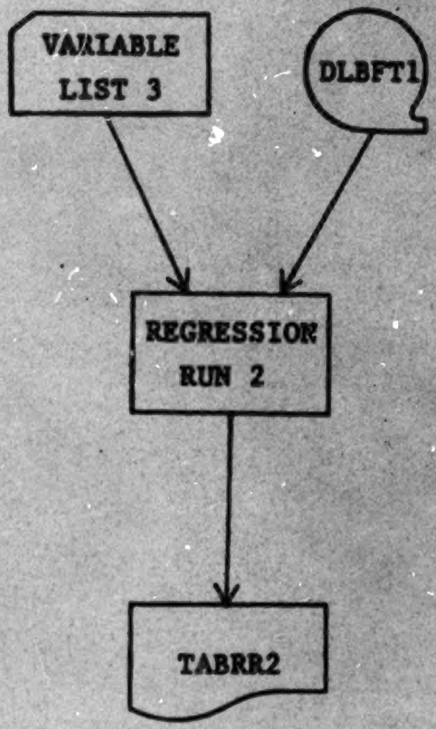
VARIABLE
LIST 3
DLBFT1
REGRESSION
RUN 2
TABRR2

Figure 18. Regression run 2.

BIBLIOGRAPHY

Aleamoni, L. M., & Graham, M. H. The relationship between CEQ ratings and instructor's rank, class size, and course level. *Journal of Educational Measurement*, 1974, *11*, 189-202.

American Psychological Association. *Publication manual of the American Psychological Association* (Rev. ed.). Washington, D.C.: Author, 1967.

American Psychological Association. *Publication manual of the American Psychological Association* (2nd ed.). Washington, D.C.: Author, 1974.

Anikeef, A. M. Factors affecting student evaluation of college faculty members. *Journal of Applied Psychology*, 1953, *37*, 458-460.

Bain, P. T. The pass-fail option: The congruence between the rationale for and student reasons in electing. *Journal of Educational Research*, 1973, *66*, 295-298.

Baird, L., & Feister, W. J. Grading standards: The relation of changes in average student ability to the average grades awarded. *American Educational Research Journal*, 1972, *9*, 431-442.

Bausell, R. B., & Magoon, J. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement, 1972, 32, 1013-1023. (a)

Bausell, R. B., & Magoon, J. The persistence of first impressions in course and instructor evaluations. Paper presented at the meeting of the American Educational Research Association, Chicago, April, 1972. (b)

Bendig, A. W. A preliminary study of the effect of academic level, sex, and course variables on the student rating of psychology instructors. Journal of Psychology, 1952, 34, 21-26.

Bendig, A. W. The relation of level of course achievement to students' instructor and course ratings in introductory psychology. Educational and Psychological Measurement, 1953, 13, 437-448. (a)

Bendig, A. W. Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. Journal of Psychology, 1953, 36, 427-433. (b)

Bendig, A. W. A factor analysis of student ratings of psychology instructors on the Purdue Scale. Journal of Educational Psychology, 1954, 45, 385-393.

Best, J. W. Research in education (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Blum, M. L. An investigation of the relation existing 'between students' grades and their ratings of the instructors' ability to teach. *Journal of Educational Psychology*, 1936, *27*, 217-221.

Brown, G. D. *System/360 job control language*. New York: Wiley, 1970.

Cal State to probe system's grading practices. *The Chronicle of Higher Education*, 1974, *8*(18), 2.

Capozza, D. R. Student evaluations, grades and learning in economics. *Western Economic Journal*, 1973, *11*, 127.

Carrier, N. A., Howard, G. S., & Miller, W. G. Course evaluation: When? *Journal of Educational Psychology*, 1974, *66*, 609-613.

Centra, J. A. Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology*, 1973, *65*, 395-401.

Centra, J. A. College teaching: Who should evaluate it? *Findings*, 1974, *1*(1), 5-8.

Committee on the Student in Higher Education. *The student in higher education*. New Haven, Connecticut: The Hazen Foundation, 1968.

Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis*. New York: Wiley, 1971.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity and usefulness. *Review of Educational Research*, 1971, *41*, 511-535.

Costin, F., & Grush, J. E. Personality correlates of teacher-student behavior in the college classroom. Journal of Educational Psychology, 1973, 65, 35-44.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Downgrading no-grade. Time, 1974, 103(5), 66.

Doyle, K. O., Jr., & Whitely, S. E. Student ratings as criteria for effective teaching. American Educational Research Journal, 1974, 11, 259-274.

Duke University Computation Center. TSAR user's manual. Durham, North Carolina: Author, 1971.

Dziuban, C. D., & Shirkey, E. C. On the psychometric assessment of correlation matrices. American Educational Research Journal, 1974, 11, 211-216.

Eble, K. E. Improving college teaching. Phi Delta Kappan, 1971, 52, 283-285. (a)

Eble, K. E. Study indicates teaching is poor. College Management, 1971, 6, 29. (b)

Eble, K. E., & The Conference on Career Development. Career development of the effective college teacher. Washington, D.C.: American Association of University Professors, 1971.

Elmore, P. B., & LaPointe, K. A. Effects of teacher sex and student sex on the evaluation of college instructors. Journal of Educational Psychology, 1974, 66, 386-389.

Etaugh, A. F. Reliability of college grades and grade point averages: Some implications for prediction of academic performance. Educational and Psychological Measurement, 1972, 32, 1045-1049.

Fahey, G. L. Student rating of teaching, some questionable assumptions. Paper presented at the conference at the University of Pittsburg Institute for Higher Education, Pittsburg, December, 1970.

Fliger, H. If Johnny gets an "F." U.S. News and World Report, 1973, 75(2), 72.

French-Lazovik, G. Predictability of students' evaluations of college teachers from component ratings. Journal of Educational Psychology, 1974, 66, 373-385.

Frey, P. W. The ongoing debate: Student evaluation of teaching. Change, 1974, 6(1), pp. 47-48; 64.

Gadzella, B. M. College students' views and ratings of an ideal professor. College and University, 1968, 44, 89-96.

Garber, H. Certain factors underlying the relationship between course grades and student judgments of college teachers (Doctoral dissertation, University of Connecticut, 1964). Dissertation Abstracts, 1964, 26, 6512. (University Microfilms No. 66-00847)

Garber, H. Some relationships between course grades and student judgments of college teachers. Paper presented at the meeting of the American Educational Research Association, New York, February, 1967.

Gessner, P. K. Evaluation of instruction. Science, 1973, 180, 566-570.

Granzin, K. L., & Painter, J. J. A new explanation for students' course evaluation tendencies. American Educational Research Journal, 1973, 10, 115-124.

Group for Human Development in Higher Education. Faculty development in a time of retrenchment. New Rochelle, New York: Change Magazine, 1974.

Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. Journal of Educational Psychology, 1936, 27, 197-216.

Hills, J. R. Consistent college grading standards through equating. Educational and Psychological Measurement, 1972, 32, 137-146.

Holmes, D. S. The relationship between expected grades and students' evaluations of their instructors. Educational and Psychological Measurement, 1971, 31, 951-957.

Holmes, D. S. Effects of grades and disconfirmed grade expectations on students' evaluations of their instructor. Journal of Educational Psychology, 1972, 63, 130-133.

Jaeger, R. M., & Freijo, T. D. Some psychometric questions in the evaluation of professors. Journal of Educational Psychology, 1974, 66, 416-423.

Kawecki, G. Teacher evaluations: Are they accurate? Connecticut Daily Campus, 1974, 77(119), pp. 1; 7.

Keefer, K. E. Characteristics of students who make accurate and inaccurate self-predictions of college achievement. *Journal of Educational Research*, 1971, *64*, 401-404.

Kerlinger, F. N. *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, 1973.

Kirschenbaum, H., Napier, R., & Simon, S. B. *Wad-ja-get? the grading game in American education*. New York: Hart, 1971.

Ladas, H. Grades: Standardizing the unstandardized standard. *Phi Delta Kappan*, 1974, *56*, 185-187.

LeComte, E. Does student power corrupt? *National Review*, 1974, *26*, 1166.

Lee, C. B. (Ed.). *Improving college teaching*. Washington, D.C.: American Council on Education, 1967.

Maeroff, G. I. Students' scores again show drop. *New York Times*, 1973, *123*(42), pp. 1; 26.

Malinowski, S. N. Not discarded. *Connecticut Daily Campus*, 1974, *77*(100), 2.

McCracken, D. D. *A guide to FORTRAN IV programming*. New York: Wiley, 1965.

McDaniel, E. D., & Feldhusen, J. F. Relationships between faculty ratings and indexes of service and scholarship. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970, 619-620. (Summary)

McKeachie, W. J. Student ratings of faculty. AAUP Bulletin, 1969, 55, 439-444.

McKeachie, W. J. Research on student ratings of teaching. Paper presented at the conference at the University of Pittsburg Institute for Higher Education, Pittsburg, December, 1970.

McKeachie, W. J. The decline and fall of the laws of learning. Educational Researcher, 1974, 3(3), 7-11.

McNemar, Q. Psychological statistics (3rd ed.). New York: Wiley, 1962.

Menzie, J. C. An analysis of the process of teacher evaluation in the community college. Los Angeles, Calif.: University of California at Los Angeles, 1973. (ERIC Document Reproduction Service No. ED 083 960)

Moellenberg, W. To grade or not to grade; is that the question? College and University, 1973, 49, 5-13.

Nie, N. H., Bent, D. H., & Hull, C. H. Statistical package for the social sciences. New York: McGraw-Hill, 1970.

Nunnaly, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Ockerman, E. W. Student input in changing grading systems. College and University, 1972, 47, 390-399.

Owen, S. V. Student ratings of teacher competence: A cautionary note. Paper presented at joint meeting of the Northeastern Educational Research Association and the National Council for Measurement in Education, Ellenville, New York, November, 1974.

Pambookian, H. S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. Journal of Educational Psychology, 1974, 66, 52-56.

Pape, R. Evaluate instructors each semester. Connecticut Daily Campus, 1974, 77(106), 2. (a)

Pape, R. Evaluation in midstream. Connecticut Daily Campus, 1974, 77(96), 2. (b)

Parent, E. R., Vaughan, C. E., & Wharton, K. A new approach to course evaluation. Journal of Higher Education, 1971, 42, 133-138.

Peck, R. Can students evaluate their education? Education Digest, 1971, 36, 29-31.

Postman, N. A D+ for Mr. Ladas. Phi Delta Kappan, 1974, 56, 187-188.

Potter, D., Nalin, P., & Lewandowski, A. The relation of student achievement and student ratings of teachers. Paper presented at the meeting of the American Educational Research Association, New Orleans, March, 1973.

Rodin, M. Students on teachers. Change, 1974, 6(2), 5.

Rodin, M., & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Rummel, R. J. Applied factor analysis. Evanston, Illinois: Northwestern University Press, 1970.

Silverberg, S. Grad finds grades inflating. Connecticut Daily Campus, 1974, 77(113), 2.

Slysz, W. D. An evaluation of statistical software in the social sciences. Communications of the ACM, 1974, 17, 326-332.

Spady, W. G. Authority, conflict, and teacher effectiveness. Educational Researcher, 1973, 2(1), 4-10.

Stallings, W. A., & Singhal, S. Some observations on the relationships between research productivity and student evaluations of courses and teaching. Paper presented at the meeting of the American Educational Research Association, Los Angeles, February, 1969.

Stanley, J. C. K-R 20 as the stepped up mean item intercorrelation. National Council on Measurements Used in Education Yearbook, 1957, 14, 78-92.

St.Onge, K. R. Let's get back to essentials. Change, 1974, 6(5), pp. 6-7; 62.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Tatsuoka, M. M., & Tiedeman, D. V. Statistics as an aspect of scientific method in research on teaching. Chapter IV in Handbook of research on teaching, N. L. Gage (Ed.), Chicago: Rand McNally & Co., 1963.

Too many A's. Time, 1974, 104(20), 106.

Touq, M. S., & Feldhusen, J. F. The relationship between students' ratings of instructors and their participation in classroom discussion. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, February, 1973.

Treffinger, D. J., & Feldhusen, J. F. Predicting students' ratings of instruction. Proceedings of the 78th Annual Convention of the American Psychological Association, 1970, 621-622. (Summary)

Turnbull, W. W. Testing shifts from 'which' to 'how.' New York Times, 1974, 123(42), 82.

Vacon, B. Faculty evaluate grade distribution. Connecticut Daily Campus, 1974, 77(106), pp. 1; 4. (a)

Vacon, B. Grades: Do they pass the test? Connecticut Daily Campus, 1974, 77(105), pp. 1; 5. (b)

Vacon, B. New requirements limit students on Dean's List. Connecticut Daily Campus, 1974, 77(73), 3. (c)

Veldman, D. J. Fortran programming for the behavioral sciences. New York: Holt, Rinehart and Winston, 1967.

Voeks, V. W., & French, G. M. Are student-ratings of teachers affected by grades? Journal of Higher Education, 1960, 31, 330-334.

Watkins, B. T. A-to-F grading system heavily favored by undergraduate, graduate institutions. The Chronicle of Higher Education, 1973, 8(8), 2.

Widlak, F. W., McDaniel, E. D., & Feldhusen, J. F. Factor analyses of an instructor rating scale. Paper presented at the meeting of the American Educational Research Association, New Orleans, February, 1973.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Worthington, L. H., & Grant, C. W. Factors of academic success: A multivariate analysis. Journal of Educational Research, 1971, 65, 7-10.

Zelby, L. W. Student-faculty evaluation. Science, 1974, 183, 1267-1270.